4-2018

# Exploiting user and venue characteristics for fine-grained tweet geolocation

Wen Haw CHONG
*Singapore Management University*, whchong.2013@phdis.smu.edu.sg

Ee Peng LIM
*Singapore Management University*, eplim@smu.edu.sg

**DOI:** https://doi.org/10.1145/3156667

Published in ACM Transactions on Information Systems, Volume 36 Issue 3, April 2018
Article No. 26 pp. 1-34.
https://doi.org/10.1145/3156667


# Exploiting User and Venue Characteristics for Fine-Grained Tweet Geolocation

WEN-HAW CHONG and EE-PENG LIM, Singapore Management University

Which venue is a tweet posted from? We call this a fine-grained geolocation problem. Given an observed tweet, the task is to infer its discrete posting venue, e.g., a specific restaurant. This recovers the venue context and differs from prior work, which geolocats tweets to location coordinates or cities/neighborhoods.

First, we conduct empirical analysis to uncover venue and user characteristics for improving geolocation. For venues, we observe *spatial homophily*, in which venues near each other have more similar tweet content (i.e., text representations) compared to venues further apart. For users, we observe that they are *spatially focused* and more likely to visit venues near their previous visits. We also find that a substantial proportion of users post one or more geocoded tweet(s), thus providing their location history data. We then propose geolocation models that exploit spatial homophily and spatial focus characteristics plus posting time information. Our models rank candidate venues of test tweets such that the actual posting venue is ranked high. To better tune model parameters, we introduce a learning-to-rank framework. Our best model significantly outperforms state-of-the-art baselines. Furthermore, we show that tweets without any location-indicative words can be geolocated meaningfully as well.

CCS Concepts: • **Information systems** → **Data mining**; *Geographic information systems*;

Additional Key Words and Phrases: Tweet geolocation, learning to rank, spatial homophily, spatial focus

## 1 INTRODUCTION

On Twitter, users post tweets from their current locations with the option of associating their tweets with location coordinates. Such geocoded tweets can be mined for insights on visit behavior or to support various applications, such as venue recommendation, disaster relief management [15], and location-based advertising. For example, for venue recommendation, knowing that a Twitter user has just tweeted from a shopping mall, the user may be recommended to visit an art gallery next to the mall. At the same time, location-based advertising can send promotion e-coupons related to the mall or art gallery to facilitate more user spending. For disaster relief management, users trapped in disasters or unforeseen incidents can potentially be geolocated through their tweets for disaster relief, information collection, or evacuation. However, studies [1, 18] have shown that as much as 98% of tweets are not geocoded. This motivates the need for location inference with tweet geolocation techniques [1, 2, 25, 26, 39].

Authors' addresses: W.-H. Chong and L. Ee-Peng, Singapore Management University, 80 Stamford Road, Singapore 178902; emails: whchong.2013@phdis.smu.edu.sg, eplim@smu.edu.sg.

**Problem.** In this work, we conduct *fine-grained geolocation*[1] [21, 25, 26, 27], which links tweets to the specific venues from which they were posted. We consider fine-grained, discrete venues[2]—e.g., restaurants, offices, and pubs—rather than coarse-grained locations such as cities or neighborhoods. We cast fine-grained geolocation as a learning-to-rank problem. Given a non-geocoded tweet from a city, we rank venues in the city such that highly ranked venues are more likely to be the posting venue. To evaluate ranking accuracy, we use a standard ranking metric, the Mean Reciprocal Rank.

Fine-grained geolocation recovers the venue context, which is useful for applications. Basically, a tweet is associated with different venue contexts when it is posted from different venues. This is true even if the candidate venues are adjacent to each other with effectively the same location coordinates. Hence, our task is very different from most of the earlier works on coarse-grained geolocation [1, 18, 20, 23]. These coarse-grained geolocation works link tweets to cities/neighborhoods or to location coordinates, which may be too coarse for many useful applications. Also note that, during geolocation, the input includes observed content from the test tweet. This differs from work on location/venue prediction [34, 46, 49], which predicts ahead of time and without observing any test content where the user will check in or visit next. To avoid confusion with such works, we have not used the term *prediction*.

**Analogy to document retrieval.** If one regards test tweets as queries and venues as documents, then fine-grained geolocation is akin to the document retrieval task [30]. However, there is one interesting difference in that venues have geographical locations and are naturally ordered in space, whereas this is not the case for documents in traditional retrieval tasks. As will be seen, the spatial positions of venues can be exploited for geolocation. In this work, each tweet is also posted from only one venue. In the analogy of document retrieval, there is only one relevant document for each query.

**Challenges.** Tweets are short and colloquial and may be posted from any one of the thousands of candidate venues in a given city or area of interest. Hence, fine-grained tweet geolocation is highly challenging. For example, a tweet "having dinner" can arise from any of the numerous food venues or even at one's home. Some prior work [25] mitigated this challenge by performing fine-grained tweet geolocation for tweets with location-indicative words only, i.e., words used mostly at very few locations, e.g., "airport." Tweets with such words are thus easier to be geolocated. Here, we geolocate both tweets with and without location-indicative words. To achieve better geolocation performance and to perform fine-grained geolocation on any tweets, we shall exploit the characteristics of users and venues, as uncovered by our empirical analysis.

For fine-grained geolocation, it is also challenging to acquire ground-truth data for meaningful experiments. Tweets have to be associated with the specific venues, instead of just the location coordinates. A popular strategy [2, 27] is to leverage location apps, such as Foursquare, in which users associate their posts with specific venues. Besides adopting this, we also propose a novel strategy of linking tweets to venues based on Foursquare users posting tweets and check-ins within a short time period (see Section 2.2).

**Empirical Analysis.** For more effective geolocation, we first study some useful characteristics of venues and users, namely, spatial homophily, spatial focus, and the availability of location history. We first exploit the venues to investigate *spatial homophily* with respect to fine-grained spatial locations. Spatial homophily is a concept that has been studied at coarse geographical resolution [1, 3]. This concept means that social media content from the same city/region is more likely to share common words than content from different cities/regions, possibly due to geographical

[1]Portions of this work appeared in [7].

[2]Such venues are also called *fine-grained locations* in [21, 26].

bias of language use in Twitter. For example, "Tube" is commonly used to refer to the subway system in London, but hardly used in a similar fashion for Singapore. *However, at a much finer spatial scale, such as between venues in a city, is spatial homophily still observable?* Our empirical studies indicated yes. Venues near each other tend to have more similar content than venues further apart in the same city. In other words, venues near each other have more similar text representations. Furthermore, spatial homophily is stronger for tweet content generated using a location app (e.g., Foursquare) than that for tweet content that is posted not using a location app. Next, we focus on the user aspect. We show that while the proportion of geocoded tweets on Twitter is small [1, 18], they are posted by a substantial proportion of users. This justifies the design of personalized models that exploit user location history in location-related applications. In addition, we show that users are *spatially focused* and are more likely to visit venues that are near each other. This characteristic can be readily incorporated into probabilistic models for geolocation.

In addition to the user and venue characteristics that surfaced through our empirical analysis, we also note that venues have the characteristic of varying popularities with different times of the day. We called this *venue temporal popularity* and shall also exploit it for modeling.

**Approach.** Drawing from the various user and venue characteristics, we then propose several probabilistic geolocation models. We formulate our models such that parameters can be easily optimized in a learning-to-rank framework. We incorporate the loss function from [8] as a proxy for the ranking metric of mean reciprocal rank along with novel adaptations to lower computation complexity.

Via extensive experiments, we show that models incorporating user and venue characteristics such as venue temporal popularity and user location history consistently outperform pure content-based approaches. We also show our models to be useful even on tweets without words that are indicative of locations. This enables us to geolocate more tweets in applications.

**Contributions.** Our contributions are listed as follows:

(1) We approach fine-grained geolocation from a learning-to-rank framework, which prior work had only sparsely explored. To obtain more data at scale for this problem, we also propose a novel strategy of linking tweets to venues based on Foursquare users posting tweets and check-ins within a short time period.
(1) We conduct empirical analysis to surface characteristics for exploitation in models. We show that spatial homophily exists at fine granularities such that venues near each other are more similar in content. We observe this effect to be stronger for tweet content generated in association with a location app.
(2) We show that 30% to 40% of users in Twitter have location history that is useful for model building. We also show that users are spatially focused in being more likely to visit venues near each other.
(3) We propose several novel models for the fine-grained geolocation problem. For selected models, we optimized their parameter by minimizing an adapted loss function in a learning-to-rank framework.
(4) Our experiments show that the various characteristics are useful for geolocation, with venue temporal popularity and user characteristics (location history and spatial focus) achieving many improvements. Depending on the dataset and metric, our best-performing model provides ranking accuracy improvement from 6% to 60% over the naïve Bayes model.

**Paper Outline.** The rest of the article is organized as follows. We first define two kinds of geocoded tweets in this study and the corresponding datasets in Section 2. We then cover the empirical study of both user and venue characteristics in Section 3, finding that spatial homophily

Table 1. Sample Shouts

| | |
|---|---|
| 1 | **Passport photo look retarded** (@ Immigration & Checkpoints Authority w/5 others) |
| 2 | **Dread dread dread work** (@ Orchard Central in Singapore) |

*Note*: Bolded portions are user-authored comments. Only this portion is used for empirical analysis and geolocation.

exists for venue representations and that user venue visits are spatially focused. Section 4 presents our proposed fine-grained geolocation models. The experiment setup and results are given in Section 5. Section 6 provides a survey of related work before we present our conclusions in Section 7.

## 2 TWEETS WITH POSTING VENUES

In this work, we require tweets with ground-truth venues for training and testing. To find them, we exploit users who are present on both Twitter and the location app Foursquare and extract two types of tweets with posting venues. The first type is Foursquare check-in comments that users broadcast to Twitter. The second type is content authored on Twitter independently of any location app, which we then associate with venues using a very stringent criterion. As to be discussed next, we apply a different preprocessing step for each type of tweet before using the data.

### 2.1 Shouts (SHT)

These are tweets pushed from Foursquare, a highly popular location-based social networking app. We follow the setup in prior work [2, 27] to construct a convenient source of tweets with ground-truth venues.

In Foursquare, users can write comments and broadcast them to Twitter while they check in to a venue. Following Foursquare terminology, we refer to such tweets as *shouts*. As shown in Table 1, a shout contains the user-authored comment, e.g., "Passport photo look retarded," plus an app-generated portion indicating the check-in venue e.g., "(@ Immigration & Checkpoints Authority w/5 others)." We discard the latter portion, which is trivial for geolocation and not meaningful for empirical analysis. Thereafter, we use only the comments for empirical analysis and geolocation.

### 2.2 Pure Tweets (TWT)

We refer to tweets that are authored by users and non-retweets as *pure tweets*. While the pure tweets are not geocoded, we find the subset of pure tweets posted by users who also performed Foursquare check-ins around the same time. Specifically, for each pure tweet from a user $u$, we link it to $u$'s check-in that is nearest in time. If the time difference is less than a specified threshold, then we assign the check-in venue as the tweet's posting venue. We use a stringent threshold of 5 minutes. This assumes that the user is tweeting from where the user checks in if both actions are within 5 minutes of each other.

To further motivate our linking process, we also note that while the current Twitter API allows users to assign location tags, the assignments are coarse grained and at the city or neighborhood level. For example, when posting a tweet, users in Singapore can select from a list of location tags: "Central Region, Singapore" and "East Region, Singapore." Such location tags are not indicative of the posting venues, however; thus, linking is still required.

### 2.3 Datasets

In this paper, "tweets" refer to both pure tweets and shouts. Where differentiation is required, we use each term explicitly, i.e., pure tweets or shouts.

We collect check-ins and pure tweets for users from Singapore (SG) and Jakarta (JKT) who utilized Foursquare. Note that Foursquare check-ins are available only if the users have broadcasted them to Twitter. For such users, we also collect their pure tweets.

We use the datasets for our empirical analysis on spatial homophily, spatially focused users, and in our geolocation experiments. For Singapore, we collected 1,190,522 Foursquare check-ins from 2014, of which 361,899 (30.4%) involve shouts. The check-ins are posted by 29,301 users over 65,701 Foursquare venues—each venue is already characterized by name, location coordinates, and functionality, e.g., "Chinese restaurant." Hence, geolocating to each Foursquare venue is equivalent to geolocating to a venue context. We refer to this dataset as **SG-SHT**. Based on the discussed process in Section 2.2, we also collected 90,250 pure tweets from 6,424 users over 12,616 venues. We designate the dataset as **SG-TWT**. For Jakarta, the **JKT-SHT** dataset comprises 177,570 check-ins for the period 2015 to mid-2016, of which 86,343 (48.6%) are shouts. The check-ins are from 12,119 users over 45,213 venues. Linking the check-ins to pure tweets, we obtain only 1,335 pure tweets (**JKT-TWT**) posted by 592 users from 886 venues. This small number is due to platform API changes made by Twitter in 2015 that affected crawling. We use JKT-TWT only as a test set.

## 3 EMPIRICAL STUDY

### 3.1 Spatial Homophily

Users in the same city/region generate more similar social media content when compared to another city/region [3, 4] due to geographical bias in language usage in Twitter. We refer to this as spatial homophily with respect to locations. Does spatial homophily exist on a much smaller spatial scale, such as between venues? To our knowledge, spatial homophily has not been studied at the venue level, thus motivating our analysis. Given venues in the same city, we compare the content of venues near each other versus venues that are far apart. If spatial homophily exists, then venues near each other should have more similar text representations. We conduct separate experiments using two different text representations of venues based on the simple bag-of-words model and a more sophisticated word-embedding technique. Both experiments indicate that spatial homophily iexists at very fine granularities. For conciseness, we discuss the bag-of-words experiment here and defer the word embedding one to Appendix B.

Table 2 presents the results. Within each dataset, we conduct two sets of analysis. In the first set (labeled "Mixed"), we compare venues near each other regardless of their functionality. In the second set, we control for functionality by comparing venues within the same category, e.g., comparing adjacent restaurants. The venue category labels are provided by Foursquare. There are ten categories based on functionality. For better representativeness, we use the two categories "Food" and "Shop," which cover more venues. This analysis allows us to evaluate spatial homophily under mixed and non-mixed functionality conditions. Our intuition is that spatial homophily should be less observable under the mixed condition.

For brevity, we describe the procedure for the Mixed analysis. If we are controlling for venue functionality, we need to repeat only the steps on venues of the targeted category. We treat each venue as a document and use its tweets to create a TFIDF vector. Let $c(w, v)$ be the frequency of word $w$ at venue $v$, let $V$ be the number of distinct venues and let $df(w)$ be the number of venues where $w$ occurs at least once. Then the $w$th dimension of $v$'s TFIDF vector is computed as $c(w, v) \log(1 + V/df(w))$. We then conduct the following:

Table 2. Average Ratio Statistic ($\bar{R}$) and Average Proportion of Venues Where Nearest Neighbors are More (or Less) Similar in Content Compared to Non-neighbors

| Dataset | Category | More similar | Less similar | Equally similar | $\overline{R}$ |
|---|---|---|---|---|---|
| SG-SHT | Mixed | 41.71% | 19.14% | 39.15% | 0.516 |
| | Food | 50.61% | 30.95% | 18.44% | 0.476 |
| | Shop | 35.72% | 21.18% | 43.10% | 0.486 |
| SG-TWT | Mixed | 36.38% | 26.26% | 37.36% | 0.461 |
| | Food | 30.67% | 25.94% | 43.39% | 0.438 |
| | Shop | 38.63% | 29.51% | 31.86% | 0.461 |
| JKT-SHT | Mixed | 29.50% | 17.09% | 53.41% | 0.470 |
| | Food | 30.52% | 23.70% | 45.78% | 0.445 |
| | Shop | 32.20% | 18.92% | 48.88% | 0.476 |

- Find $k$ venues nearest to $v$ that are also below distance threshold $D$. This forms $v$'s nearest-neighbor set, denoted as $nb(v)$. If there are $l < k$ venues below distance threshold, $nb(v)$ will only include $l$ venues.
- Compute average cosine similarity between $v$ and nearest neighbors, denoted as $\overline{cos}_{nb}(v)$.
- Randomly sample $k$ venues more than distance $D$ away as non-neighbors, denoted as $nnb(v)$.
- Compute $\overline{cos}_{nnb}(v)$, the average cosine similarity between $v$ and non-neighbors.
- Compute the average distance from $v$ to $nb(v)$: $\overline{dist}_{nb}(v) = \frac{1}{|nb(v)|}\sum_{v' \in nb(v)} d(v, v')$, where $d(v, v')$ is the distance between $v$ and $v'$. Also, compute $\overline{dist}_{nnb}(v)$, the average distance from $v$ to $nnb(v)$.

Since Singapore and Jakarta are dense cities, we use $k = 5$ and $D = 500m$. After iterating over all venues with content, we tabulate the proportion of venues whose nearest neighbors are more similar than the non-neighbors, i.e., $\overline{cos}_{nb}(v) > \overline{cos}_{nnb}(v)$ and the proportion of venues whose nearest neighbors are less similar than non-neighbors. Since the non-neighbors are sampled randomly, we conduct 10 runs per city and average the proportions.

For each venue in each run, we also compare the cosine similarities of neighbors and non-neighbors with the following **ratio statistic**:

$$R(v) = \exp\left(\frac{-\overline{cos}_{nnb}(v)}{\overline{cos}_{nb}(v)}\right), \tag{1}$$

where the exponential function avoids computation error caused by dividing by zero. $R(v)$ is larger in terms of content; $v$ has less similar non-neighbors than neighbors. For each run, we average $R(v)$ over venues to obtain the **average ratio statistic** $\overline{R}$.

Table 2 displays the average ratio statistics and the averaged proportions. Venues with identical $\overline{cos}_{nb}(.)$, $\overline{cos}_{nnb}(.)$ fall under the "Equally similar" column in the table. These identical value cases involve venues with no common words, i.e., $\overline{cos}_{nb}(v) = \overline{cos}_{nnb}(v) = 0$. Other venues fall under the "More similar" or "L similar" column. Table 2 shows that proportions in the "More similar" column are consistently higher for all datasets than the "Less similar" column. This implies spatial homophily since venues are more similar to their neighbors than to random non-neighbors. For example, in SG-SHT, on average, 50.61% of food venues are more similar to their food venue neighbors while 30.95% are less similar when compared against non-neighbors of the food category. The difference between these two proportions is greater for SG-SHT than SG-TWT, suggesting that the spatial homophily effect is stronger for shouts than pure tweets.

Table 3. Venues (in Brackets <>) Near Each Other and Sample Shouts Demonstrating Spatial Homophily

| | |
|---|---|
| M1 | <Cha Cha Cha Mexican Restaurant & Bar><br>"Hehe finally satisfied ma Mexican food craving w momsie" |
| M2 | <El Patio Mexican Restaurant & Wine Bar><br>"Mexican Hogmany food with @joanniewalker" |
| N1 | <Executive Cafe><br>"Hotpot at NTU. Yum <3 with Lem" |
| N2 | <McDonald's><br>"At NTU's North Spine." |

Refer to the $\overline{R}$ values in Table 2. If there is no difference in cosine similarities between neighbors and non-neighbors, then Equation (1) indicates that $\overline{R}$ is expected to be $\exp(-1) = 0.368$. As can be seen, all values are higher than this. On average, a venue's non-neighbors is less similar in content than neighbors. This again indicates spatial homophily. $\overline{R}$ is also higher for SG-SHT than SG-TWT across all categories. This reaffirms that spatial homophily is stronger for shouts. One possible explanation is that, for pure tweets, users tend to share more diverse topics, which can be quite unrelated to their current venues. Different from pure tweets, shouts are authored by users as they check in to some venues. They then broadcast their shouts to Twitter, intentionally sharing their venues. Thus, users may be more likely to mention aspects related to current venues or the local area. This also implies that pure tweets are harder to geolocate compared to shouts.

Interestingly, the "Mixed" experiment, which does not control for venue functionality, exhibits spatial homophily effects that are comparable to "Food" and "Shop." On inspection, we observed various contributing factors. While moving around adjacent venues of different functionalities, users may mention local spatial characteristics, events, or may be using unique words, e.g., mentions of friends.

Table 3 illustrates examples of spatial homophily. Shouts M1 and M2 are from Mexican restaurants near each other. User mentions of Mexican food contribute to content similarity between venues. For shouts N1 and N2, they are posted from venues in Nanyang Technological University (NTU), a university in Singapore. Thus, NTU constitutes a local spatial feature and its mentions increase content similarity between venues on campus.

### 3.2 Location History

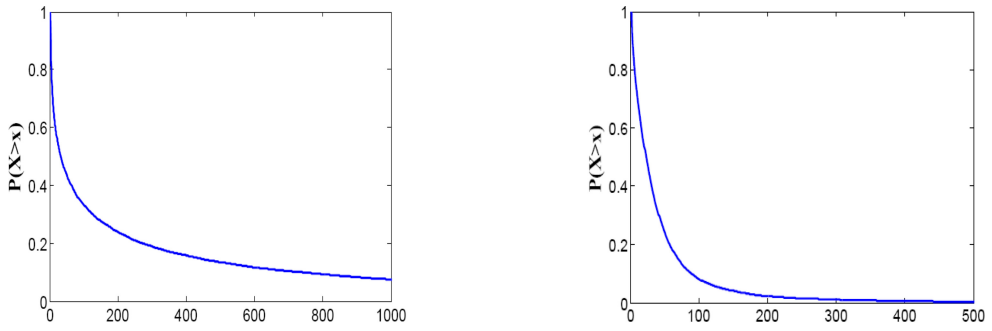
As the proportion of geocoded tweets is small, one may easily assume that they are contributed by an equally small proportion of users. For such users, the geocoded tweets constitute a personal location history that can be used to build more accurate models to geolocate their non-geocoded tweets. However, are such models widely applicable to users? We therefore need to investigate the proportion of users with personal location history.

For the purpose of this empirical analysis, we sample users independently of the datasets discussed earlier in Section 2.3. We randomly sample 50,000 Twitter users from Singapore for 2014 and from Jakarta for June to December 2016. The only sampling condition is that each sampled user has posted at least one tweet during the study period. Sampled users may or may not be active on Foursquare. Table 4 shows the statistics compiled.

As expected from prior work [1, 18], the proportion of geocoded tweets is tiny at 3.22% for Singapore and 4.62% for Jakarta. However, we find that the proportion of users posting geocoded tweets is substantial. For ease of discussion, denote the set of users who posted at least one geocoded

Table 4. Statistics for 50,000 Sampled Users from Singapore (2014) and from Jakarta (June to December, 2016)

| | Singapore | Jakarta |
|---|---|---|
| Total Tweets | 136,548,216 | 20,466,019 |
| Geocoded Tweets | 4,394,378 (3.22%) | 946,432 (4.62%) |
| Users with geocoded tweets, $\{u\}_g$ | 15,169 (30.34%) | 20,982 (41.97%) |
| Average geocoded tweets/user in $\{u\}_g$ | 289.69 | 45.11 |
| Average non-geocoded tweets/user in $\{u\}_g$ | 4532.98 | 157.48 |



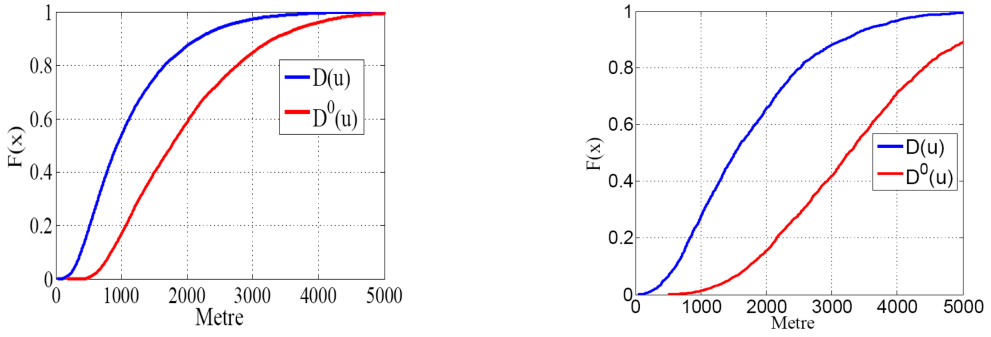

(a) Singapore (b) Jakarta

Fig. 1. CCDF for users in $\{u\}_g$. X-axis = number of geocoded tweets per user.

tweet as $\{u\}_g$. Table 4 shows that, in Singapore, $\{u\}_g$ comprises 30.34% of the sampled users. This is much larger than the value of 3.22% if one does a naïve inference based on the fraction of geocoded tweets. Similarly, in Jakarta, $\{u\}_g$ is substantial at 41.97% of the users. Such proportion characteristics arise because users in $\{u\}_g$ post both geocoded and non-geocoded tweets, with the latter at much larger counts. The last two rows of Table 4 illustrate this. On average, a Singapore user in $\{u\}_g$ posts 289.69 geocoded tweets and 4532.98 non-geocoded tweets. A similar bias in tweeting behavior can be observed for Jakarta.

Intuitively, an average user is constrained by geographical, social, or personal factors. This leads to venue revisits or the conduct of many activities (e.g., work) in geographically localized regions. Now, consider a user in $\{u\}_g$. The user has geocoded tweets with location coordinates. This location history may provide useful information on the user's visit routines and activity regions. We can then build a personalized model of the user that better geolocates the individual's other non-geocoded tweets. Obviously, this also requires sufficient geocoded tweets per user, thus motivating our next analysis.

For users in $\{u\}_g$, we examine their distribution of geocoded tweets. This gives a sense of the proportion of users with sufficient location history for learning a model. Figure 1 displays the Complementary Cumulative Distribution (CCDF) plot. The plots show that many users in $\{u\}_g$ have an adequate number of geocoded tweets. For example, Figure 1(a) indicates that, for Singapore,




(a) Singapore (b) Jakarta

Fig. 2. CDF of distance statistics of users (blue) versus null model (red). (X axis = distance in metres).

around 40% of the users in the $\{u\}_g$. set has more than 50 geocoded tweets over a 1-year period. For Jakarta, over a half-year period, the corresponding proportion is around 25%.

### 3.3 Spatially Focused Users

We say that a user is *spatially focused* if more likely to visit venues that are near the user's other visited venues. For each user, we compute a distance-based statistic to quantify the extent of spatial focus. We compare this against the expected distance statistic when a user visits the same number of venues in a random manner. We term the latter as the null model. We conduct our analysis on users with geocoded tweets tied to Foursquare (datasets SG-SHT and JKT-SHT).

Denote $\mathbb{V}_u$ as the set of venues visited by user $u$. We iterate through each venue in $\mathbb{V}_u$ and compute the distance to the nearest neighboring venue. This is averaged over all venues in $\mathbb{V}_u$. If the distance statistic is small relative to the null model (to be defined), then there is stronger evidence of spatial focus. Formally, the distance statistic is

$$D(u) = \frac{1}{|\mathbb{V}_u|} \sum_{v \in \mathbb{V}_u} \min_{v' \in \mathbb{V}_u \setminus v} d(v, v'), \tag{2}$$

where $d(,)$ measures spatial distance. $D(u)$ is easy to compute. It neither assumes any parametric form for the spatial distribution nor knowledge of the number of spatial clusters.

The null model computes the expected distance statistic if the user is not spatially focused but rather visiting venues at random. For the null model, we reassign each unique visit of user $u$ to a random venue and obtain a random venue set $\mathbb{V}_u^0$ of the same size as $\mathbb{V}_u$. We then apply Equation (2) again to compute the distance statistic $D^0(u)$. Note that, to ascertain the presence of spatial focus, it is important to compare $D(u)$ versus the null model rather than just examining its actual value. The reason is that $D(u)$ can be small even if a user is not spatially focused. For example, assume a huge geographical area containing many points that equally split the area. Let these points correspond to the coordinates visited by user $u$. When the number of points is sufficiently large, then $D(u)$ is small, although $u$ is not spatially focused. However, in this case, if we apply the null model, $D^0(u)$ will be small as well and close to $D(u)$. Thus, by comparing both values, we can avoid drawing the wrong conclusion that $u$ is spatially focused.

Figure 2 plots the Cumulative Distribution Function (CDF) of the distance statistics for Singapore and Jakarta. For each city, there is clear evidence that users are spatially focused. The red curve for the null model statistic consistently lies to the right of the blue curve for the user statistic. This implies that venues visited by users are spatially nearer each other than random. For example,

Figure 2(a) shows that if users are visiting venues randomly (red curve), then we expect only 60% to have a distance statistic of 2,000 metres or less. However, the actual behavior (blue curve) indicates that the corresponding proportion is around 90%. For Jakarta in Figure 2(b), 65% of users (blue) have a distance statistic of 2,000 metres or less, much higher than the expected proportion of 15% based on the null model (red).

**Remarks.** In short, our empirical analysis highlights that users with geocoded tweets form a significant group in Twitter, much more than what one would expect from the proportion of geocoded tweets. We also observe strong evidence that users tend to visit venues that are spatially near each other. These motivate the design of personalized models based on users' location history.

### 3.4 Venue Temporal Popularity

Each tweet is associated with a posting time, which provides a modeling linkage to venue temporal popularities. Intuitively, different venues are more popular at different times of the day, e.g., dining venues are more popular at meal times while nightlife venues are more popular at late hours. This directly affects the probability that a venue is the posting venue of a given tweet at different times of the day.

Venue temporal popularities were studied extensively in prior work [29, 34, 49]. Also, tweet posting time is always observed, in contrast to user location history. Hence, we omit empirical analysis and coverage studies. Instead, we capture the discussed intuitions by including tweet posting time for modeling. This improves geolocation performance significantly, as will be discussed in the experiment results.

## 4 MODELS

We first describe a Naïve Bayes model as the baseline model for geolocation. We then propose several probabilistic models that draw on the empirical analysis findings. We elaborate the associated notations in an inline manner for ease of reading.

### 4.1 Naïve Bayes (NB)

We denote the naïve Bayes model from [23, 25] as NB. This models the tweet content associated with each venue as a bag of words $\mathbf{w}$. Let $W$ be the vocabulary size of tweet words. We use $c(w, v)$ as the frequency of word $w$ at venue $v$ and $c(\cdot, v)$ to denote $\sum_w c(w, v)$. Given a tweet, we then rank venues by the venue probability $p(v|\mathbf{w}) \propto p(v) \prod_{w \in \mathbf{w}} p(w|v)$. The probability of word given venue $p(w|v)$ is

$$p(w|v) = \frac{c(w, v) + \alpha}{c(\cdot, v) + W\alpha}, \tag{3}$$

where $\alpha$ is the smoothing parameter that can be tuned or set at 1 for Laplace smoothing.

### 4.2 Spatial Smoothing (NB+S)

Our earlier empirical analysis had demonstrated the presence of spatial homophily where venues near each other are more similar in content. To consider this effect, we propose adding spatial smoothing to the naïve Bayes model NB. For each word $w$ at the ego venue $v$, we extend the definition of $p(w|v)$ with word frequencies of $v$'s set of spatial neighbors, denoted by $nb(v)$. The spatially smoothed $p(w|v)$ is defined as:

$$p(w|v) = \frac{c(w, v) + \alpha + \frac{\gamma}{|nb(v)|} \sum_{v_i \in nb(v)} c(w, v_i)}{c(\cdot, v) + W\alpha + \frac{\gamma}{|nb(v)|} \sum_{v_i \in nb(v)} c(\cdot, v_i)}, \tag{4}$$

where $0 \leq \gamma \leq 1$ is the weight factor. By setting $\gamma$, we adjust the spatial smoothing strength on word frequencies from $v$'s neighbors. When $\gamma = 1$, a word $w$ found in every $v$'s neighbor will be equivalent to a single $w$ occurrence in $v$. Otherwise, the words from neighbors are weighted less than the native words in $v$. Also recall that our earlier analysis shows that spatial homophily exists even without controlling for venue functionalities. Thus, we do not need to restrict neighbors to be of the same category as the ego venue.

### 4.3 Tweet Posting Time (NB+S+T)

The previous models mainly exploit the tweet content. As tweet content is short, ranking accuracy may be low due to information sparsity. We thus explore user and/or venue characteristics to improve performance. As previously mentioned, the posting time of tweets is readily available. This ties to the characteristic that certain venues are more popular at different times of the day, making them more likely to be the posting venues of tweets. Hence, given a tweet posted at time of day $t$, we incorporate time into the model as follows:

$$p(v|\mathbf{w}, t) \propto p(v|t) \prod_{w \in \mathbf{w}} p(w|v), \tag{5}$$

where $p(v|t)$ accounts for venue popularity at time of day $t$.

A simple approach to compute $p(v|t)$ is to discretize $t$ into time bins—e.g., hourly—and estimate the venue distribution for every bin. However, there are boundary effects that are counterintuitive. For example, consider discretizing by hourly bins, each bin starting on the hour. Then, $t = 2{,}359$ hours and $t = 0001$ hours are in different bins although they are only 2 minutes apart. In contrast, $t = 0001$ hours and $t = 0059$ hours are 58 minutes apart but in the same bin.

Instead of binning, we model time of day $t$ as a continuous variable, which is more intuitive. We estimate $p(v|t)$ in an approach motivated by kernel density estimation (KDE) [28]. For time of day $t$, define a time interval of length $T(t)$ that covers $t$. Denote by $f(v, t)$ the number of user visits to venue $v$ in the interval $T(t)$ and let $f(\cdot, t) = \sum_v f(v, t)$. Given a test tweet with time of day $t$, we compute that

$$p(v|t) = \frac{f(v, t) + \beta}{f(\cdot, t) + V\beta}, \tag{6}$$

where $V$ is the number of distinct venues and $\beta$ is the smoothing parameter. $\beta$ can be tuned or learned (see Section 4.5).

Defining a time interval and counting the venues within is similar to applying a uniform kernel in KDE. The time interval length $T(t)$ is analogous to the kernel bandwidth. Instead of adopting a fixed interval length, we use *adaptive bandwidth selection* [28]. Basically, given a test tweet posted at time of day $t$, we search for the $k$ training tweets closest in time of day to define the time interval, i.e., $f(\cdot, t) = k$. To do this efficiently, we use a k-d tree structure [14]. Given a set of training tweets $\mathbb{T}$, insertion and search using the k-d tree has average complexity of $O(\log |\mathbb{T}|)$. We index all training tweets after converting their posting times to 2-dimensional Cartesian coordinates. We use a dimension of 2, which is the lowest possible dimension to capture the cyclical property of time of day, e.g., a quarter past noon is closer to midnight than a quarter to noon. Formally, let time of day $t$ be represented as the number of seconds past midnight. We compute the corresponding Cartesian coordinate $(t_x, t_y)$ as

$$\begin{aligned} t_x &= \sin(t/3600) \\ t_y &= \cos(t/3600) \end{aligned}. \tag{7}$$

Following Equation (7), we can readily apply k-d trees and Euclidean distance to facilitate $k$ nearest-neighbor computation given any time-of-day query. Note that the k-d tree is built on the training set and accessed during test time.

Using the parameter $k$, adaptive bandwidth selection is able to adjust the time interval length locally based on data density. Basically, during timings with sparse training points (e.g., midnight), the interval length is longer to cover $k$ nearest-neighboring training tweets, while during timings with dense training points (e.g., dinner), the interval length is shorter. This is also intuitive from the Bayesian point of view. Consider Equation (6), where $f(v,t)$ and $f(\cdot,t)$ are actual observations while $\beta$ and $V\beta$ are pseudo-observations. In fixing $f(\cdot,t)=k$, we effectively use $k$ to control the relative importance of actual and pseudo-observations to be consistent across all test tweets.

### 4.4 User Location History (NB+S+T+U)

Earlier, we showed that a substantial proportion of users have location history in the form of geocoded tweets. On average, such users also post many non-geocoded tweets, which may be targeted for geolocation. Here, we use location history to build models that are personalized to each user.

Consider the previous model NB+S+T. From Equation (5), this model can be interpreted as a Bayesian network in which the time-of-day node generates the venue node, which then generates the words. We now let the venue node generate the user node as well. Thus, we now define

$$p(v|\mathbf{w},t,u) \propto p(v|t)p(u|v)\prod_{w\in\mathbf{w}} p(w|v). \tag{8}$$

Since location history is specific to users, it is more intuitive to compute $p(v|u)$ instead of $p(u|v)$. $p(v|u)$ can also be represented by two-dimensional distributions over geographical space, which is convenient for interpretation and visualization. By the property $p(u|v)=p(v|u)p(u)/p(v)$ and assuming constant $p(u), p(v)$, we have $p(u|v)\propto p(v|u)$. Thus, the probability term $p(u|v)$ in Equation (8) can be replaced by $p(v|u)$.

To model $p(v|u)$, recall that the spatial focus property means users are more likely to visit venues spatially near previously visited venues. To capture this idea, we extend the distance statistic from Equation (2) and define $p(v|u)$ as

$$p(v|u) \propto \exp\left(-S\cdot \min_{v'\in\mathbb{V}_u} d(v,v')\right), \tag{9}$$

where $\mathbb{V}_u$ is defined previously as the set of venues in $u$'s location history and $S\geq 0$ is the decay parameter that will be learned (see Section 4.5). A large $S$ means that $p(v|u)$ decreases faster with increasing distance between $v$ and the nearest venue in $\mathbb{V}_u$. Equivalently, we are making the model more sensitive to the spatial focus property. In contrast, if $S=0$, we disregard the spatial focus property. In our experiments, the learned optimal $S$ varies across datasets. On average, $S=0.116$ for SG-SHT, $S=0.025$ for JKT-SHT, and $S=0.302$ for SG-TWT.

Equation (9) defines an affinity vector over venues, specific to user $u$, whereby $p(v|u)$ are the vector elements. This vector is fixed if user $u$'s location history is not updated. Thus, one can precompute the affinity vectors for users to geolocate their tweets more efficiently. Last, note that for notation simplicity, we have defined Equation (9) in terms of distances between venues. In fact, it is not required for the specific venues to be known in the location history. It suffices for only the location coordinates of geocoded tweets to be known. Thus, the proposed model is applicable to more users, including those whose geocoded tweets are not associated with specific venues.

**Query Likelihood Model.** Equation (8) can be interpreted as a query likelihood model (see Section 12.2 of [30]) in the framework of traditional document retrieval. In the query likelihood

model, the probability of document $\mathbf{d}$ given query $q$ is computed as $p(\mathbf{d}|q) \propto p(q|\mathbf{d})p(\mathbf{d})$. In Equation (8), venues are analogous to documents while the test tweet's user and content comprises a query with accompanying meta-information. The posting time is used to assign a non-uniform prior to the venues (i.e., documents).

### 4.5 Learning to Rank

Given a tweet, one desires its posting venue to be ranked high. Thus, there is only one relevant venue and the Mean Reciprocal Rank (MRR) is a suitable metric. Consider a set of tweets $\mathbb{T}$. Let the ranked position of the $i$th tweet's posting venue $v_i$ be $r_i$, where $0 \leq r_i \leq V - 1$. The MRR with respect to tweet set $\mathbb{T}$ is defined as

$$MRR(\mathbb{T}) = \frac{1}{|\mathbb{T}|} \sum_{i=1}^{|\mathbb{T}|} \frac{1}{1 + r_i}, \tag{10}$$

which is the average of the reciprocal ranks for each tweet in $\mathbb{T}$. The highest possible value for $MRR(\mathbb{T})$ is 1.0, whereby each tweet's posting venue is ranked at the top. $MRR(\mathbb{T})$'s lowest possible value is $1/V$, whereby each tweet's posting venue is ranked at the bottom.

We can optimize parameters of our models with respect to MRR via tuning or Learning to Rank (LTR). For models with few parameters, e.g., NB and NB+S, tuning can be done with grid search over the parameter space in order to maximize MRR directly. However, for more complicated models with more parameters, tuning cost increases at an exponential rate. In contrast, LTR requires a proxy function in place of MRR and may be susceptible to local optima. However, LTR can utilize gradient information for more fine-grained optimization and scales better with increasing model parameters. Considering the computation cost of tuning versus LTR and the number of model parameters, we apply different approaches to different models. For NB and NB+S, we adopt tuning based on grid search. For NB+S+T and NB+S+T+U, we adopt LTR. To further motivate our choice of using LTR instead of tuning, assume that each parameter is tuned over a grid of $\tau$ values. Then, for NB+S+T+U, which has 4 parameters, tuning requires applying the model on the tuning set for a total of $\tau^4$ times. This is much more expensive than, for example, tuning for NB+S, which only requires applying the model for $\tau^2$ times.

LTR requires one to define an appropriate objective function. First, Equation (10) can be re-expressed as:

$$MRR(\mathbb{T}) = \frac{1}{|\mathbb{T}|} \sum_{i=1}^{|\mathbb{T}|} \left[ 1 + \sum_{v \neq v_i} \mathbf{I}(p_\Theta(v) > p_\Theta(v_i)) \right]^{-1}, \tag{11}$$

where $\mathbf{I}(.)$ is the indicator function and $p_\Theta(v_i)$ is the probability of the posting venue $v_i$ for the $i$th tweet computed from some geolocation model $\Theta$ (e.g., NB+S). Thus, maximizing MRR is equivalent to maximizing some function constructed from multiple 0-1 loss functions. Note that the indicator function has a gradient of 0 except at the point of discontinuity, where the gradient is ill defined. Hence, it is infeasible to maximize MRR directly [8] via LTR. Instead, one approximates MRR maximization by minimizing a proxy loss function, whereby a good proxy should approximate the 0-1 loss well while retaining a sufficient gradient for learning. Various loss functions are possible, e.g., logistic loss. However, in recent work, [8] has proposed the log-log loss function as a better alternative to logistic loss. This motivates us to introduce the log-log loss function into our models that have been selected for LTR. For the selected models, we construct the loss function over venue pairs for minimization.

*4.5.1 Loss Function.* For a posting venue $v_i$ to be ranked high, $p_\Theta(v_i)$ should be large while $p_\Theta(v)$ should be small for $v \neq v_i$, i.e., non-posting venues. For computation convenience, we use

log probabilities for ranking. Let $z_\Theta(v_i, v) = \ln p_\Theta(v_i) - \ln p_\Theta(v)$. The log-log loss function for a tweet with posting venue $v_i$ is

$$L_\Theta(v_i) = \sum_{v \neq v_i} \ln(1 + \ln(1 + \exp(-z_\Theta(v_i, v)))) = \sum_{v \neq v_i} \ln(1 + R_\Theta(v_i, v)), \tag{12}$$

where $R_\Theta(v_i, v) = \ln(1 + \exp(-z_\Theta(v_i, v)))$. To obtain the global loss function, one computes and sums Equation (12) over the set of tweets considered:

$$G_\Theta(\mathbb{T}) = \sum_{i=1}^{|\mathbb{T}|} L_\Theta(v_i). \tag{13}$$

*4.5.2 Reparameterization.* With the loss function defined, we can perform gradient descent to minimize it. However, there are constraints on the parameters. The smoothing parameters $\alpha$, $\beta$, and $S$ are required to be non-negative. The spatial weight factor $\gamma$ has to satisfy the constraint $0 \leq \gamma \leq 1$. Instead of constrained optimization, we incorporate the above constraints by reparameterizing the model as follows:

$$\begin{aligned} \alpha &= x_\alpha^2 \\ \gamma &= (1 + \exp(-x_\gamma))^{-1} \\ \beta &= x_\beta^2 \\ S &= x_S^2 \end{aligned}, \tag{14}$$

where $x_\alpha$, $x_\beta$, $x_S$, and $x_\gamma$ are the new set of parameters. These can now be easily learned from unconstrained optimization.

*4.5.3 Gradients.* We minimize the loss function via stochastic gradient descent. Here, we illustrate deriving the gradient for one parameter $x_S$ for one model: NB+S+T+U model. For notation brevity, let $\Theta$ represent NB+S+T+U. By chain rule,

$$\frac{\partial L_\Theta(v_i)}{\partial x_S} = \sum_{v \neq v_i} \frac{\partial \ln(1 + R_\Theta(v_i, v))}{\partial R_\Theta(v_i, v)} \frac{\partial R_\Theta(v_i, v)}{\partial z_\Theta(v_i, v)} \frac{\partial z_\Theta(v_i, v)}{\partial x_S}. \tag{15}$$

For NB+S+T+U, we have that $p_\Theta(v) = p(v|\mathbf{w}, t, u)$, thus:

$$\frac{\partial z_\Theta(v_i, v)}{\partial x_S} = \frac{\partial \ln p(v_i|\mathbf{w}, t, u)}{\partial x_S} - \frac{\partial \ln p(v|\mathbf{w}, t, u)}{\partial x_S}, \tag{16}$$

where

$$\frac{\partial \ln p(v|\mathbf{w}, t, u)}{\partial x_S} \propto \frac{\partial \ln p(u|v)}{\partial x_S} = -2x_S \cdot \min_{v' \in \mathbb{V}_u} (d(v, v')) \tag{17}$$

and $\partial \ln p(v_i|\mathbf{w}, t, u)/\partial x_S$ is computed similarly. The gradients for other model parameters are illustrated in Appendix A. The gradients for the model NB+S+T are derived in a similar fashion.

*4.5.4 Complexity Reduction.* Let $\mathbb{T}$ be the set of training tweets and let $V$ be the number of distinct venues. For each training tweet, there is one posting venue and $V$-1 non-posting venues. Consequently, for each training tweet, if we construct pairwise loss between the posting venue and all other non-posting venues, then there are $V$-1 pairs. The overall computational complexity for training is then $O(|\mathbb{T}|V)$.

We can reduce the complexity by reducing the number of pairs considered per training tweet. The simplest approach is to randomly sample $M$ proportion of pairs per training tweet (e.g., $M = 0.25$) such that $MV < V$-1. On top of this random sampling scheme, we propose further adaptations to reduce the complexity while enabling changes in the loss function to be more correlated to changes in MRR. We achieve this by assigning greater weights to training tweets that contribute

more to MRR. Such tweets already have their posting venues ranked high and are intuitively more important. For example, assume two tweets at the start of training: tweet 1 with its venue ranked at position 0, i.e., $r_1 = 0$, and tweet 2 with its venue ranked at position 99, i.e., $r_2 = 99$. The overall MRR is $(\frac{1}{0+1} + \frac{1}{99+1})/2 = 0.505$, with a contribution of 0.5 from tweet 1 and 0.005 from tweet 2. Tweet 1 is thus much more important than tweet 2. As training proceeds, model parameters evolve and may lead to changes in the venue rankings of both tweets. However, any changes in the rank of tweet 1's venue will affect MRR much more than tweet 2.

The loss function as defined by Equations (12) and (13) do not reflect the varied importance of training tweets. Furthermore, given some reduction in the loss, not all reciprocal ranks associated with test tweets are simultaneously improved. Instead, there is a mixture of improvement, decline, or no change. Continuing from the earlier example, it is plausible for a given loss reduction to improve the ranking of tweet 2's venue to position 49, while tweet 1's associated ranking may drop to position 1. This leads to a reduced MRR of $(\frac{1}{1+1} + \frac{1}{49+1})/2 = 0.26$, even though the loss has decreased. Hence, to better correlate loss reduction with MRR improvement, it is important to improve or maintain the ranking accuracy for tweets already associated with high reciprocal rank. To achieve better correlation, we let more important training tweets contribute more pairs. Specifically for the $i$th tweet at the start of the training phase, we construct the pairwise loss to $M_i(V-1)$ other venues where $M_i$ is a proportion computed as

$$M_i = \frac{M}{1 + \exp(-1/r_{i,0})}, \tag{18}$$

where $r_{i,0}$ is the ranked position of the posting venue for the $i$th tweet at the start of training, i.e., the 0th iteration. Tweets contribute more pairs (are assigned more importance) based on their associated reciprocal rank such that $M_i = M$ for $r_{i,0} = 0$ and is close to $0.5M$ for large values of $r_{i,0}$. For example, a tweet with its posting venue perfectly ranked at the start of training contributes ($V$-1)$M$ pairs to the global loss function while a tweet with a very poorly ranked venue contributes close to only 0.5($V$-1)$M$ pairs. The computational complexity is now $O(V \sum_i^{|\mathbb{T}|} M_i)$. Except for extreme and unlikely cases in which all posting venues are perfectly ranked at the start of training, the new computational complexity is lower than $O(|\mathbb{T}|V)$, enabling training to be conducted faster.

**Other Sampling Strategies.** It is possible to formulate other sampling strategies to exploit various types of information associated with the training set, such as user visit history and venue categories. For example, given a training tweet posted by user $u$ from venue $v$, one may want to sample non-posting venues in a manner specific to $u$, e.g., sampling equally from venues visited/not visited by $u$. Alternatively, the sampling process can be specific to the posting venue $v$. For example, one can sample non-posting venues belonging to the same functional category (e.g., restaurants) as $v$. This means that posting and non-posting venues are less differentiated and may make parameter learning more sensitive to generate distinct ranks for such venues. In short, various strategies can be studied to understand their strengths and weaknesses. We defer such exploration to future work.

## 5 EXPERIMENTS

### 5.1 Setup

We conduct fine-grained geolocation experiments to achieve the following:

(1) Compare our models with each other and other state-of-the-art baselines.
(2) Assess the importance of incorporating various user and venue characteristics, such as user location history and temporal venue popularity.

We split the datasets SG-SHT, SG-TWT, and JKT-SHT into training, tuning, and test sets. Model parameters are learned from the training set to minimize the loss on the tuning set. We include venues as ranking candidates only if they have at least 5 tweets in the training set. We also filter out stop words and rare words (frequency <4). The test set consists of test cases of tweets, each posted from some venue by a user with location history. On inspection, we noticed "easy" test cases, in which a user repeatedly uses a highly unique word with each post from a certain venue. This makes the unique word highly indicative of the posting venue, leading to high ranking accuracy for such cases. To make the problem more challenging, we filter them from the training set as follows: for each test case with user $u$ and posting venue $v$, we exclude $u$'s other tweets posted at $v$ from the training set. In other words during training, applied models do not observe any postings of $u$ from venue $v$.

For each dataset, we conduct 20 runs in which, for each run, we sample 5,000 tweets for testing/tuning and use the remainder for training. From the sampled set, we use 1,000 tweets for tuning and the remainder for testing. Due to various types of filtering discussed above, the number of test cases per run is less than 4,000. The average numbers of test cases are reported with the results for each experiment.

### 5.2 Models Applied

We compare the following models:

- KL: This model [27] assigns scores to venues based on posting time information, e.g., hour of day, and the Kullback-Leibler divergences between the smoothed language models of tweets and venues. The KL divergences are transformed and linearly combined with the venue probabilities to form ranking scores.
- TFIDF: We represent venues and tweets as TFIDF vectors in terms of content. Given a test tweet, we use cosine similarity to retrieve and rank venues. This is very similar to the method in [20].
- GMM: This models [3] each word as a Gaussian mixture over 2-d space and a test tweet as the product of Gaussian mixtures. Venues are ranked by the probability that the product of Gaussian mixtures generates their coordinates. Since words that are indicative of spatial regions should have relatively few numbers of modes, we follow [3] and set the number of clusters to 3.
- VDOC: The topic model VDOC in [6] models the generation of check-ins and venue-related comments in the form of Foursquare tips. By treating tweets as tips and ignoring the check-in mode, we extend VDOC to model tweet generation. To generate each tweet, the venue first generates the topic. The topic then generates the posting user and the tweet words. In our experiments, we used 40 topics after observing that this is sufficient for optimal ranking performance.
- KDE: KDE [19] integrates kernel density smoothing into multinomial naïve Bayes to geolocate tweets to grid cells. Given cell $c$, geolocation is based on the probability $p(c)\prod_{w\in\mathbf{w}} p(w|c)$, whereby $p(c)$ and $p(w|c)$ are smoothed using Gaussian kernels. To apply the method for geolocating to venues, we extend it to compute $p(v|c)p(c)\prod_{w\in\mathbf{w}} p(w|c)$. Given venue $v$ located in cell $c$, $p(v|c)$ is estimated by counting tweets posted from venue $v$ over all tweets posted within cell $c$. We experiment with grid sizes of 1 km and 500 m and report results from the latter due to its better performance.
- NB: This is the naïve Bayes, content-only approach from [23, 25]. We observed better performance with uniform venue probabilities, i.e., $p(v|\mathbf{w}) \propto \prod_{w\in\mathbf{w}} p(w|v)$, and report the associated results. We tune the smoothing parameter $\alpha$ using grid search: $\alpha$ is varied from

Table 5. Average MRR for SG-SHT

| Model | MRR | Improvement over NB |
|---|---|---|
| KL | 0.04021 | −58.045% |
| TFIDF | 0.03571 | −62.740% |
| GMM | 0.02495 | −73.967% |
| VDOC | 0.03683 | −61.571% |
| KDE | 0.06655 | −30.561% |
| NB | 0.09584 | 0% |
| NB+S | 0.09620 | 0.376% |
| NB+S+T | 0.09966 | 3.986% |
| NB+S+T+U | 0.10271 | 7.168% |

*Note*: On average, there are 2,626.2 test cases and 10,814.5 venues to rank per run.

0.1 to 1.5 in steps of 0.1. The value associated with the optimal tuning set MRR is then selected.

- NB+S: This extends the NB model with spatial smoothing. For spatial smoothing, we use $k = 5$ nearest neighbors of each venue to smooth the word probabilities. We tune the smoothing parameter $\alpha$ and the spatial weight factor $\gamma$ using grid search over $(\alpha, \gamma)$ values from (0.1,0.0) to (1.5,1.0), at intervals of 0.1.
- NB+S+T: This uses content with spatial smoothing plus tweet posting time, which relates to venue temporal popularity. For parameter learning, we apply LTR with mini-batch stochastic gradient descent. We set $M$ (see Equation (18)) at 0.25. We use 15 epochs and 50 mini-batches, each mini-batch consisting of 20 tweets. To account for local optimal, we randomly initialize and train 5 instances per model. We then select the instance with the highest tuning set MRR to apply on the test set.
- NB+S+T+U: This uses content with spatial smoothing, posting time, and user location history, thus exploiting all user and venue characteristics. We apply LTR with the same set-up as described above.

### 5.3 Results on Shouts

In the first experiment, we train and test models on the datasets SG-SHT and JKT-SHT. Tables 5 and 6 present results for Singapore and Jakarta shouts, respectively. Note that MRR figures are not directly comparable across datasets since we are ranking with different venue sets. For JKT-SHT, there are also fewer venues to rank, making it easier to achieve high MRR.

In both Tables 5 and 6, KL, TFIDF, GMM, VDOC, and KDE substantially underperform the NB model. KL includes posting time information but fails to outperform NB anyway. Evidently, modeling each shout with a smoothed language model, as done by KL, is inadequate. This, in turn, affects the computing of KL divergences between the word distributions of shouts and venues. TFIDF also consistently has low MRR partly because it is not optimized for ranking. Also, if a test shout and a posting venue share no common word, cosine similarity is 0 and the venue will be ranked low. This may be overly stringent. The topic model VDOC performs poorly despite its model complexity. This may be due to the fact that model parameters are optimized with respect to the formation of coherent topics rather than with respect to MRR. For GMM, performance is poor as we have to geolocate even shouts where words do not have peaky Gaussian distributions. Among the approaches that are inferior to NB, KDE is the best performing. Primarily, this approach models and

Table 6. Average MRR for JKT-SHT

| Model | MRR | Improvement over NB |
|---|---|---|
| KL | 0.03019 | −77.667% |
| TFIDF | 0.04193 | −68.982% |
| GMM | 0.09767 | −27.748% |
| VDOC | 0.05849 | −56.732% |
| KDE | 0.10665 | −21.105% |
| NB | 0.13518 | 0% |
| NB+S | 0.13623 | 0.777% |
| NB+S+T | 0.14618 | 8.137% |
| NB+S+T+U | 0.14824 | 9.661% |

*Note*: On average, there are 975.9 test cases and 2,713.75 venues to rank per run.

smooths the word distributions of grid cells instead of venues. Thus, word distributions are learned at a coarser level and are suboptimal for fine-grained geolocation.

Both Tables 5 and 6 exhibit similar trends from the NB model onwards. MRR improves as we add spatial smoothing and additional characteristics to the models. For adjacent models—e.g., NB versus NB+S—we have also conducted significance testing with the Wilcoxon signed rank test. The differences between all models are statistically significant at $p$ value of 0.05.

Comparing NB and NB+S, spatial smoothing improves MRR slightly, which can be attributed to the presence of spatial homophily. The improvement is small but consistent across different runs. This may be due to the limited strength of spatial homophily at fine granularities. We also note that prior work on coarse-grained geolocation [4] had reported limited improvement from spatial smoothing even when using location-indicative words only. For example, in [4], which geolocates users' cities with accuracy as the metric, the improvement from spatial smoothing is less than 1%. We also reason that even without smoothing, we are already capturing much of the spatial homophily effect. Recall that this means that venues near each other have more similar content. In the NB model, we are modeling the venue content directly anyway, thus implicitly accounting for spatial homophily in a downstream manner.

For both cities, substantial improvement comes from exploiting temporal venue popularity and location history. For example, NB+S+T provides 3.986% improvement over NB in Table 5. For Jakarta in Table 6, the corresponding improvement is 8.137%. Thus, venue popularity with the time of day plays a role. Adding user location history helps to increase MRR even more, with NB+S+T+U being consistently the best performing model in both tables. This shows that location history is highly useful. Also, recall that our modeling approach captures the idea that users are spatially focused in being more likely to visit venues that are near each other. The experiment results further validates this.

### 5.4 Results on Pure Tweets

In this experiment, we train and test our models on pure tweets from Singapore (SG-TWT). Results are displayed in Table 7. We only rank venues appearing in pure tweets. This results in an average of 2,783.55 venues to rank per run. Also, JKT-TWT has too few pure tweets for training; thus, we do not use it in this experiment.

The trend in Table 7 is mostly similar to that of the previous experiment on shouts. KL, TFIDF, GMM, and VDOC are poor performers. KDE performs better than these techniques, but loses out

Table 7. Average MRR for SG-TWT

| Model | MRR | Improvement over NB |
|---|---|---|
| KL | 0.03790 | −33.310% |
| TFIDF | 0.02059 | −63.769% |
| GMM | 0.01385 | −75.629% |
| VDOC | 0.01986 | −65.054% |
| KDE | 0.05349 | −5.877% |
| NB | 0.05683 | 0% |
| NB+S | 0.05718 | 0.612% |
| NB+S+T | 0.07600 | 33.526% |
| NB+S+T+U | 0.09229 | 62.015% |

*Note*: On average, there are 2,061.9 test cases and 2,783.55 venues to rank per run.

slightly to NB. Spatial smoothing again provides only slight improvement over the NB model, although it is statistically significant over 20 paired runs. The exploitation of venue temporal popularity and user location history provides very sharp improvement. NB+S+T+U again has the highest MRR, with over 60% improvement from NB.

Typically, MRR is not compared across experiments that rank different numbers of items. However, here, we can make certain statements by comparing Tables 7 and 5. In Table 7, for pure tweets, we rank fewer venues, but obtain mostly lower MRR than Table 5 for shouts. Since we have fewer venues to rank, the task should have been easier, resulting in a higher MRR. The lower MRR thus implies that it is more challenging to rank venues for pure tweets than shouts. This observation is also consistent with our empirical analysis (Table 2), whereby we have observed spatial homophily to be stronger for shouts than pure tweets. Also, pure tweets may be about more diverse topics not related to the posting venue. Obviously, this will impact ranking accuracy.

If the contents of pure tweets are not highly indicative of venues, then characteristics such as temporal venue popularity and user location history become relatively more important. This is illustrated by the huge gains in MRR as we move from model NB to NB+S+T/NB+S+T+U. The percentage improvement is much larger in Table 7 than the case for shouts in Table 5.

### 5.5 Applying Shout Models to Pure Tweets

In this experiment, we explore whether models that are trained to rank using shouts (i.e., model NB and extensions) will perform well on pure tweets. The motivation is that, in applications, it is easier to form training sets using shouts that are already associated with venues than tweets that require labeling or some linking process. We apply the models trained on SG-SHT to test tweets from SG-TWT. We also train models with JKT-SHT and test on JKT-TWT. For test cases, we use pure tweets that contain one or more words from the shout content vocabulary. We use the set of shout venues for ranking. This makes it possible to compare with the results for shouts.

Tables 8 and 9 depict the respective results for Singapore and Jakarta. The trend is similar to training/testing with pure tweets or shouts. Spatial smoothing contributes a small improvement while substantial improvements occur as we model additional characteristics. Clearly, temporal venue popularity and location history remain highly important.

For each city, we cross-compare the results for pure tweets and shouts, i.e. Tables 8 versus 5 and Tables 9 versus 6. Clearly, MRR is consistently lower for pure tweets across all models. This affirms again that pure tweets are more challenging to geolocate than shouts. This is so even though we

Table 8. Average MRR from Applying SG-SHT Models to Test on SG-TWT

| Model | MRR | Improvement over NB |
|---|---|---|
| NB | 0.04021 | 0% |
| NB+S | 0.04028 | 0.1741% |
| NB+S+T | 0.04993 | 24.173% |
| NB+S+T+U | 0.05821 | 44.765% |

*Note*: On average, there are 31,946.2 test cases and 10,814.5 venues to rank per run.

Table 9. Average MRR from Applying JKT-SHT Models to Test on JKT-TWT

| Model | MRR | Improvement over NB |
|---|---|---|
| NB | 0.10571 | 0% |
| NB+S | 0.10596 | 0.237% |
| NB+S+T | 0.14043 | 32.845% |
| NB+S+T+U | 0.14241 | 34.718% |

*Note*: On average, there are 363.15 test cases and 2,713.75 venues to rank per run.

are using pure tweets from users who also posted shouts. This should limit the differences in topics and vocabulary.

### 5.6 Stratified Experiment

Finally, we compare geolocation for tweets with and without Location-Indicative (LI) words. We also examine whether we can obtain meaningful geolocation accuracy for the latter. LI words suggest a venue or spatial region with high probability, e.g., "airport." Typically, ignoring tweets without LI words can improve performance [3, 4] for the task of inferring a user's home location. This is because users typically post multiple tweets, some of which are more informative of their home location. However, we have a different task of geolocating individual tweets. Tweets without LI words were considered not appropriate for fine-grained geolocation and excluded in an earlier work [25]. Equivalently, they were regarded as noise. Depending on the strictness of the criteria for detecting LI words, a substantial fraction of data may be discarded. This is undesirable in applications.

We adopt the approach in [25] to detect LI words. LI words have high occurrence probability in at least one venue and occur at relatively few venues. Words are scored based on the TFIDF measure as follows:

$$LI(w) = \max_{v} \left\{ p(w|v) \log \left( \frac{V}{df(w)} \right) \right\}. \tag{19}$$

We point out that Equation (19) encapsulates some word popularity effects due to the term $p(w|v)$. Thus, more popular words tend to have higher scores, although this is offset to some extent by the lower inverse document frequency inherent in such words. Empirically, we observe a larger fraction of tweets indicated as containing LI words compared to other word scoring measures [3]. In [25], they applied the NB model after using Equation (19) to filter out tweets with no LI words. Here, we conduct more extensive experiments by stratifying tweets based on whether they contain LI words or not, followed by applying our proposed models on both types of tweets.

Table 10. Results for Stratified Experiment

| Dataset | Statistics | Models | $MRR(\mathbb{L})$ | $MRR(\neg\mathbb{L})$ |
|---|---|---|---|---|
| SG-SHT | $\lvert\mathbb{L}\rvert$=1726.5<br>$\lvert\neg\mathbb{L}\rvert$=899.7<br>$V$=10814.5 | NB | 0.11748 | 0.05441 |
| | | NB+S | 0.11755 | 0.05529 |
| | | NB+S+T | 0.11841 | 0.06376 |
| | | NB+S+T+U | 0.12184 | 0.06608 |
| | | Random | 9.123E-4 | |
| SG-TWT | $\lvert\mathbb{L}\rvert$=484.65<br>$\lvert\neg\mathbb{L}\rvert$=1577.25<br>$V$=2783.55 | NB | 0.11270 | 0.03983 |
| | | NB+S | 0.11352 | 0.04007 |
| | | NB+S+T | 0.12154 | 0.06204 |
| | | NB+S+T+U | 0.13441 | 0.07939 |
| | | Random | 3.057E-3 | |
| JKT-SHT | $\lvert\mathbb{L}\rvert$=464.05<br>$\lvert\neg\mathbb{L}\rvert$=511.85<br>$V$=2713.75 | NB | 0.22806 | 0.05125 |
| | | NB+S | 0.22954 | 0.05195 |
| | | NB+S+T | 0.23153 | 0.06912 |
| | | NB+S+T+U | 0.23279 | 0.07191 |
| | | Random | 3.126-3 | |
| JKT-TWT | $\lvert\mathbb{L}\rvert$=137.25<br>$\lvert\neg\mathbb{L}\rvert$=225.9<br>$V$=2713.75<br>(Based on JKT-SHT venues) | NB | 0.19240 | 0.05279 |
| | | NB+S | 0.19285 | 0.05288 |
| | | NB+S+T | 0.20687 | 0.09996 |
| | | NB+S+T+U | 0.20609 | 0.10354 |
| | | Random | 3.126-3 | |

*Note*: $\mathbb{L}$ and $\neg\mathbb{L}$ are, respectively, the set of test tweets with and without LI words, with associated mean reciprocal rank of $MRR(\mathbb{L})$ and $MRR(\neg\mathbb{L})$. The model "Random" denotes a random ranking model. Statistics and results shown are averaged over 20 runs.

If tweets without LI words are not meaningful for geolocation, then when geolocating such tweets, the expected ranking performance is equal to geolocating random noise. This means that the ranking of candidate venues is random, with uniform probabilities over all reciprocal rank outcomes. The expected Reciprocal Rank ($RR$) from random ranking can then be computed as

$$\mathrm{E}_{\mathrm{Random}}[RR] = \sum_{i=1}^{V} p\left(\frac{1}{i}\right)\frac{1}{i} = \frac{1}{V}\sum_{i=1}^{V}\frac{1}{i} \tag{20}$$

where "Random" is a model that does random ranking. The expected MRR then follows by averaging over the number of geolocated tweets. Subsequently, for tweets without LI words, we shall compare each model's MRR against the expected MRR from random ranking.

Equation (19) results in LI scores that are dataset dependent, e.g., $V$ varies across different datasets. Instead of specifying dataset-dependent thresholds, we simply designate the top 5% scoring words as LI words for each dataset. Our experiment setup is similar to that in Section 5.1, except that test tweets are now stratified into tweets with LI words (denote as set $\mathbb{L}$) and tweets without LI words ($\neg\mathbb{L}$). We compute MRR for each set of test tweets, i.e., $MRR(\mathbb{L})$ and $MRR(\neg\mathbb{L})$.

Table 10 displays the results of the stratified experiment for all datasets. Also included in the table is the expected MRR for the model "Random," which randomly ranks candidate venues. This regards the tweets as noise independently of whether they contain LI words or not. Hence, MRR values are equal across both tweet sets $\mathbb{L}$ and $\neg\mathbb{L}$.

As shown in Table 10, it is easier to geolocate tweets with LI words than tweets without. Consistently across all models for all datasets, $MRR(\mathbb{L})$ is substantially higher than $MRR(\neg\mathbb{L})$. For both

MRR values, there is also an improving trend as we incorporate more characteristics into the models. From the trend corresponding to $MRR(\mathbb{L})$, it is clear that even if we adopt the filtering process of [25] and focus only on geolocating tweets from $\mathbb{L}$, our proposed approaches provide consistent improvements.

It is important to note that Table 10 shows that $MRR(\neg\mathbb{L})$ for various models is orders of magnitude higher than the random baseline (model Random). For example, in SG-SHT, the model NB+S+T+U gives an MRR of 0.06608, which is 72.43 times that of 9.123E-4 from random ranking. This implies that we are achieving meaningful geolocation accuracy even for tweets without LI words. Second, for tweet set $\neg\mathbb{L}$, there is consistent improvement in geolocation accuracy attained from our models. Hence, there is useful information that can be progressively incorporated to geolocate such "noisy" tweets. Thus, it may not be necessary to discard such tweets, as advocated in [25].

### 5.7 Performance Analysis

From the earlier results in Tables 5, 6, and 7, we have seen that NB+S+T+U outperforms NB+S+T. Comparing both models, the difference in average MRR is small for SG-SHT and JKT-SHT at 0.00305 and 0.00206, respectively, while for SG-TWT, the difference is larger at 0.01629. Although NB+S+T+U achieves only a small increase in MRR for some datasets, the improvement is consistent across multiple runs for all datasets and has high statistical significance ($p$ value $< 0.01$). Thus, location history does provide some useful information for geolocation, which motivates the analysis in this section.

The goal of the current analysis is to examine how the performance gains attained by NB+S+T+U over NB+S+T vary with the amount of users' location history. To this end, we quantify location history with two criteria: the number of distinct venues that a user had visited (i.e., posted tweets from) and the number of visits that the user accumulated over all venues. For each dataset with multiple runs (SG-SHT, JKT-SHT, and SG-TWT), we accumulate test tweets over 10 runs and group them into four bins of equal sizes based on their users' location history, i.e., the first bin corresponds to users with the least history while the last bin corresponds to users with the most history. Since we used four bins, the bins are also referred to as quartiles; we use both terms interchangeably.

For each test tweet, we subtract the reciprocal rank attained by NB+S+T from that obtained from NB+S+T+U. This difference is then averaged over all test tweets within each quartile. Figure 3 plots the MRR differences for each dataset based on the two binning criteria of distinct venues and visit counts. In each figure in Figure 3, numbers below each bin indicate the range of location history covered. Also, ties have to be distributed between bins such that the bins are equal-sized. For example, the leftmost bin of Figure 3(a) covers test tweets whose users have distinct venues ranging from 1 to 34 in their location history. Users of test tweets in the second bin have distinct venues ranging from 34 to 74. Thus, some users in these two bins share the same distinct venue count of 34.

Across all quartiles for both binning criteria, NB+S+T+U provides gains in MRR over NB+S+T. This is consistent across the three datasets. However, the extent of improvement varies across different quartiles. A pattern emerges whereby the largest MRR gains are usually attained over the second and/or third bin from the left. Equivalently, improvement is largest for users with a moderate amount of location history compared to users with less or more location history. For example, in Figure 3(f), which corresponds to SG-TWT, MRR gains are largest for the middle two bins, i.e., tweets from users with visit counts ranging from 13 to 75. Tweets from users with less ($\leq$13) or more ($\geq$75) visits in their location history experience less improvement.

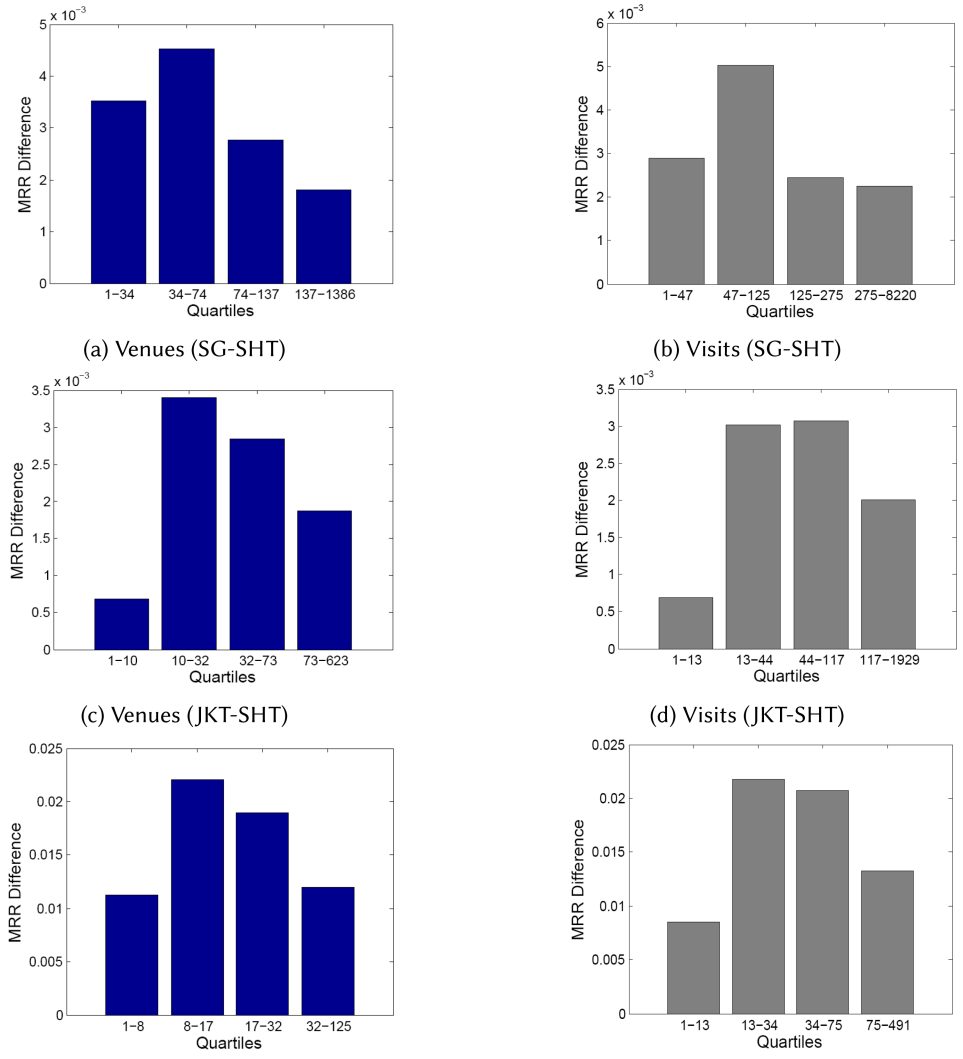



Fig. 3. Average differences in MRR between models NB+S+T+U and NB+S+T. Higher bars means that NB+S+T+U attained larger improvement over NB+S+T. Test tweets are divided into bins/quartiles based on the number of distinct venues ("Venues") and the number of visits ("Visits") in their users' location history. The number of binned tweets are 25,898 for SG-SHT, 9,429 for JKT-SHT and 19,978 for SG-TWT. For Figures (a), (c), and (e), labels on the X axis represent the range of distinct venues covered by each bin. For Figures (b), (d), and (f), X-axis labels are the range of visit counts covered by each bin.

Clearly, sparse location history limits the extent of improvement that NB+S+T+U can make. However, it can be unintuitive that gains are not monotonic with respect to the amount of location history. One reason is that user behavior is confounded with the amount of location history such that users with more location history are also visiting more venues all over the city and exhibiting a long-tailed effect. This may cancel out some of the benefits derived from more location history. For example, it is more difficult to geolocate tweets for a user who spreads visits over hundreds of

Table 11. Sample Test Tweets from SG-SHT to Illustrate Improvement of NB+S+T Over NB+S

| ID | Time of day | <Posting venue>:Tweet content | $\Delta RR$ | Ranked position (NB+S) | Ranked position (NB+S+T) |
|---|---|---|---|---|---|
| S1 | 16:10:53 | <Ion Orchard>: "**Remind** me to never **step** into **ion** on a **Sunday**.." | 0.667 | 2 | 0 |
| S2 | 18:15:59 | <Golden Village (Yishun)>: "**White House** Down!" | 0.4 | 9 | 1 |

*Note*: For each tweet, bolded words are words used for geolocation, i.e., after filtering off stop-words and rare words. $\Delta RR$ is the difference in reciprocal rank of the posting venue when applying NB+S+T versus NB+S. The last two columns show the ranked position of posting venues obtained under each model (in brackets). Note that the best possible ranked position is 0, corresponding to a reciprocal rank of 1. See Equation (10).

venues compare to another user who is mainly focused on a few dozen venues. In separate studies, we have measured the entropy of the users' distributions over venues. This is found to be higher for users with a higher number of visits in their location history. Consistent with this, we also found the number of visits to be highly correlated with the number of distinct venues, with the Pearson's correlation exceeding 0.85 across all three datasets. Thus, users with higher visit counts are also spreading their visits more widely over different venues, possibly making their tweets harder to geolocate.

### 5.8 Case Studies

In this section, we first illustrate examples in which sample tweets are geolocated more accurately from the inclusion of temporal venue popularity and user location history for modeling. We then examine cases in which the inclusion of location history does not provide improvements. This motivates the case for future work.

*5.8.1 Temporal Venue Popularity.* In Table 11, we compare sample tweets geolocated using the models NB+S and NB+S+T. Tweet S1's posting venue is a popular shopping mall in Singapore, <Ion Orchard>. Based on the venue probabilities from model NB+S, the posting venue is placed at position 2 (i.e., $r_{S1} = 2$), behind two other venues, both of which are Catholic churches. This can be explained by the fact that tweets posted from churches often contain the term "Sunday" due to Sunday services. However, with the posting time of 16:10:53, i.e., a Sunday afternoon, it is more probable for the tweet to be posted from the mall rather than from churches. This is because malls tend to be more popular than churches on Sunday afternoons. NB+S+T is able to exploit this additional information and assigns higher probability to <Ion Orchard>, making the posting venue the top ranked. The change in reciprocal rank is thus $\Delta RR = \frac{1}{(0+1)} - \frac{1}{(2+1)} = 0.667$.

For S2, the tweet was posted from <Golden Village (Yishun)>, a movie theatre. In this case, the tweet mentioned a movie title and is indicative of movie theatres. Hence, for both geolocation models, the top ranking candidate venues for the tweet are all movie theatres. However, even in this case, posting time information is still useful since the movie theatres differ in popularities based on the time of day. This may be due to differences in the screening schedule across different theatres. With the exploitation of temporal venue popularity, NB+S+T ranks the actual posting venue at position 1, an improvement of 8 places over that achieved by NB+S.

*5.8.2 Location History.* Table 12 lists three sample tweets that have been geolocated using the models NB+S+T and NB+S+T+U. Recall that the latter model assumes that each user is more likely

Table 12. Sample Test Tweets from SG-SHT to Illustrate Improvement of NB+S+T+U Over NB+S+T

| ID | Dist. to nearest user venue (m) | <Posting venue>: Tweet content | $\Delta RR$ | Ranked position (NB+S+T) | Ranked position (NB+S+T+U) |
|---|---|---|---|---|---|
| S3 | 42.2 | <Woodlands Regional Bus Interchange>: "**Hahaha** 168 **bus ride** with **mah homie** - with Eezah" | 0.056 | 8 | 5 |
| S4 | 49.5 | <Manna Story>: "**Korean food @OldLadyFang**" | 0.3 | 4 | 1 |
| S5 | 956.8 | <Republic Polytechnic>: "8am **class**" | 0.076 | 14 | 6 |

*Note*: Here, $\Delta RR$ is the difference in reciprocal rank of the posting venue when applying NB+S+T+U versus NB+S+T. The second column shows the distance of the posting venue to the next nearest venue visited by the same user.

to post from candidate venues near one's other visited venues. Thus, for each tweet, we also list the distance from the posting venue to the nearest venue in the posting user's training venues (second column of Table 12). Also, recall in our experiment setup that each user's set of training venues specifically excludes posting venues of the user's test tweets.

Tweet S3 is posted from a bus station <Woodlands Regional Bus Interchange>. For S3's user, that user's nearest venue in the training set is 42.2 m away. This turns out to be a subway station <Woodlands MRT Station>. While S3's content is indicative of a bus-related venue, there are many such venues (e.g., bus stops, bus interchanges) in Singapore. With the tweet content and spatial smoothing, NB+S only manages to rank the posting venue at position 8. By further exploiting a user's location history, NB+S+T+U geolocates S3 with higher accuracy, ranking the posting venue at position 5. This example is intuitive as well for Singapore since many commuters have to transfer between subways and buses when commuting. Thus, both subway and bus stations are frequently co-visited. S4 is posted from a Korean restaurant <Manna Story>. The user's nearest training venue is just 49.5 m away, which we observed to be a Starbucks cafe. The user is conducting activities such as dining and drinking at venues around the same area. For S5, the user's nearest training venue is 956.8 m away, which is a library. In this case, there is ranking improvement even though the nearest venue is relatively far from the posting venue, as compared to the previous two examples. Hence, the spatial focus property may still be applicable even if posting venues are sparsely distributed over space.

*5.8.3 Negative Cases.* To motivate further research, we examine negative cases in which NB+S+T+U performs worse than NB+S+T. Table 13 lists three such test tweets. Tweet S6 is mainly written in Malay and posted from a theme park <Universal Studios Singapore>. The user is not spatially focused around S6's posting venue, with the nearest venue in the user's location history being the airport at around 21 km away. On investigation, we also found that the user has extremely sparse location history, with the airport constituting the only training venue. This makes it difficult for NB+S+T+U to exploit location history. Compared to the model NB+S+T, the rank of posting venue returned by NB+S+T+U is lower, i.e., rank 3. Nevertheless, NB+S+T+U still ranks the posting venue reasonably high since it also exploits other information, such as tweet content and time. In particular, the words "transformer" and "mummy" refer to rides at <Universal Studios Singapore> and are indicative of the theme park. Hence, although there are numerous other candidate venues nearer to the airport, they are not scored higher than the posting venue.

S7 is posted from a border crossing west of Singapore. The user is not spatially focused around this venue with the user's nearest training venue at around 22 km away. In contrast to S6's user,

Table 13. Sample Test Tweets Where NB+S+T+U Results in Lower Rank Positions of Posting Venues Compared With Those Returned by NB+S+T

| ID | Dist. to nearest user venue (m) | <Posting venue>:Tweet content | $\Delta RR$ | Ranked position (NB+S+T) | Ranked position (NB+S+T+U) |
|---|---|---|---|---|---|
| S6 | 21,663.6 | <Universal Studios Singapore>: "**Transformer ama mummy nya keren** parah. **Mau lagi**." | −0.083 | 2 | 3 |
| S7 | 21,875.0 | <Tuas Checkpoint (Second Link)>: "Off to " | −0.293 | 2 | 24 |
| S8 | 2727.7 | <Ikea>: "**Meatballs** for " | −0.0571 | 4 | 6 |

*Note*: Here, $\Delta RR$ is the difference in reciprocal rank of the posting venue when applying NB+S+T+U versus NB+S+T.

S7's user has substantial location history. However, S7's user mostly visits venues in the central and northern part of Singapore, far from S7's posting venue. Thus, there is some deviation by the user from the user's usual activity area. In this case, NB+S+T+U returns a lower rank for the posting venue at position 24 while some venues from the central and north of Singapore are ranked higher.

Finally, S8 is posted from <Ikea> with the nearest user venue at about 2.7 km away, which is a less drastic case than S6 and S7. This user's location history has a good number of visits; however, the user is more active in the central business and shopping area of Singapore rather than the suburb area where <Ikea> is located. Hence, there is insufficient spatial focus around <Ikea> for NB+S+T+U to better geolocate S8.

In short, the cases discussed here highlight scenarios in which NB+S+T+U may be inadequate and are grounds for future work. S6 pertains to users with sparse location history, which may be common for tourists or new users and is akin to the cold start problem. A possible mitigation for this is to include geometric weights in the NB+S+T+U model (Equation (8)) such that the relative importance between tweet content, posting time, and location history can be tailored to each user. For new users with little location history, the latter can be assigned smaller importance. S7 and S8 pertain to users who deviate significantly from their usual visitation behavior. This can be due to users seeking novelty [50] and visiting new venues or users changing their visitation behavior over time. The latter can be for various reasons, e.g., change of workplace and moving to different housing. For better geolocation, it will be interesting in future work to incorporate the aspects of novelty seeking and behavior evolution into our models.

## 6 RELATED WORK

We discuss empirical analysis conducted in prior work that motivates our own studies. This is followed by a survey of prior work in coarse-grained and fine-grained geolocation.

### 6.1 Spatial Homophily

Spatial homophily with respect to locations was not explicitly studied, although some geographical topic modeling work [1, 12, 18, 47] implied spatial homophily at a coarse spatial level. Ahmed et al. [1] proposed a hierarchical topic model that automatically infers both the hierarchical structure over content and over the size and position of geographical locations. In the topic hierarchy, topics at a higher level correspond to broad regions whereas topics at a lower level correspond to more fine-grained locations, e.g., a neighborhood. Hong et al. [18] proposed an approach that

models content in tweets based on topical influence, user's interest and geographical influence. Geographical influence affects tweet contents, causing the probability of certain words to deviate from a global background word distribution. Yin et al. [47] used tags from geocoded Flickr images to infer region-specific topics whereby words close in space are more likely to belong to the same region and are more likely to be clustered into the same topic. We also note the work by Eisenstein et al. [12], who proposed a multilevel generative model based on cascading topic models. Their model recovers coherent topics and their regional variants while identifying geographic areas of linguistic consistency. In short, the above cited works imply the presence of geographical topics or geographically influenced content on a coarse spatial level.

Some other works [9, 13, 48] implicitly assumed spatial homophily at a more fine-grained neighborhood level. Mobility patterns and venue features are used to infer neighborhoods of various functionalities or characteristics within a city, e.g., a shopping or residential neighborhood or neighborhoods with different demographics. Cranshaw et al. [9] clustered venues in a city based on both spatial proximity and social affinity. The latter is based on representing each venue as a bag of check-in users. They show that distinctive clusters arise, representing neighborhoods of different characteristics. Falher et al. [13] characterized neighborhoods using features derived from check-ins at neighborhood venues. They also explored finding neighborhoods of similar functions across different cities using the earth-mover's distance as the metric. Yuan et al. [48] infer the functions of neighborhoods with the Dirichlet Multinomial Regression [33] topic model. They regard neighborhoods as documents, venue information as metadata, and human mobility patterns from taxi rides as words. In summary, neighborhoods are clusters of venues having similar functions or characteristics; thus, within the same neighborhood, venues should have more similar content, as suggested by spatial homophily.

### 6.2 Spatial Focus

Our notion of spatially focused users can be related to more restrictive user mobility patterns, namely, proximity of visits to home and proximity between consecutive visits.

*6.2.1 Visitation Proximity to Home.* Users are more likely to visit venues near their home locations. Pontes et al. [38] studied the relationship between home locations and mobility patterns on a coarse spatial scale. They analyzed user activities in Foursquare that are indicative of mobility patterns, e.g., tips (comments about visited venues) and venue mayorships (most frequent visitor). They found that users tend to engage in such activities at their residing cities and that they frequently revisit venues. Cho et al. [5] utilize check-ins and cellphone logs to show that users focus their visits around individual activity centers, such as the home or workplace. This supports their formulation of a visitation model based on Gaussian Mixtures. They also found that users revisit venues with substantial probability. Doan and Lim [10] conducted analysis at more fine-grained spatial resolution, within individual cities. They obtained the exact home coordinates of users by exploiting check-ins with indicative comments, e.g.,"Home sweet home!". Regarding these users, they showed that the check-in probabilities decrease for venues with increasing distances from users' home locations. Other works [37, 43] implicitly exploit the idea that user visits are spatially concentrated near their home locations. The work in [37] used majority voting and mean statistics on geocoded visit data. Tasse et al. [43] recursively partition space into grids of uniform cells and then find the mode, i.e., the cell with the most number of check-ins. By repeating this process recursively, they are able to infer the home location.

*6.2.2 Proximity Between Consecutive Visitations.* This means that consecutive venue visits over time tend to be close by. Thus, given a user's current venue, he is more likely to next visit nearby venues than venues further away. Noulas et al. [34] showed that the probability distribution of

spatial distance between consecutive check-ins exhibits a decreasing trend that resembles an inverse power law. Shorter distances are more likely to appear than longer distances, although the latter still has small, non-negligible probabilities. The study in [35] used the complementary cumulative distribution function on inter-check-in distances and arrived at very similar findings. There is also concurrence with the finding in [41] that human walk patterns exhibit statistically similar features as Levy walks [44]. The study was of very high resolution, conducted using mobility track logs from participants carrying GPS receivers. It was found that people tend to visit nearby places and occasionally distant places. In another work, Yuan et al. [49] studied Gowalla and Foursquare check-ins to uncover a similar characteristic, which they called *spatial influence*.

*6.2.3 Remarks.* If users tend to visit venues near their home, then by the transitivity property, users are also more likely to visit venues near any of their previously visited venues, i.e., they are spatially focused users. Considering proximity between consecutive visits [34, 35, 41, 49] and the observation that users revisit venues [5] or activity regions, we can arrive at a similar characteristic. Thus, one can regard spatial focus as a much more general characteristic that is applicable even if one has no knowledge of a user's home location or current location.

### 6.3 Coarse-Grained Geolocation

We review coarse-grained geolocation, as it is a well-studied research topic related to fine-grained geolocation. Coarse-grained geolocation seeks to geolocate tweets or users at the city or region level. There are two different tasks, as discussed next.

*6.3.1 User Geolocation.* The first task infers the home city or region of users by exploiting the content over multiple tweets posted by each user. For this, Cheng et al. [4] modeled the distribution of words (collected globally from multiple users) over space, such that LI words can be identified from model parameters. The idea is that such words should have high local focus and a fast dispersion, i.e., (1) it is very frequent at some central spatial point and (2) usage rapidly declines as one moves away from the central point. One can then use LI words found in the tweets of users to infer their home locations. Chang et al. [3] also exploited LI words. However, to detect such words, they applied Gaussian Mixture Models (GMM) instead. Words with probability mass that are spatially focused on a small area are then picked out as LI words. Han et. al. [16] compared various approaches: statistical methods, e.g., hypothesis testing; information theory, e.g., word entropy; and heuristics-based approaches, e.g., TFIDF to identify LI words. They found that geolocation performance of the various methods varies greatly with the number of top-ranked words. Jurgens [22] geolocated users based only on their social relationships, independent of any tweet content. The idea is to spatially propagate location assignments through the social network, using only a small number of initial locations. This assumes that users are likely to be near their friends. With the same intuition, Rahimi et al. [40] employed spatial propagation over friendship networks constructed from user mentions in tweets. They further incorporated text-based geolocation priors into their network, showing that this joint exploitation of text and social network information performs better than text-only and network-only approaches.

*6.3.2 Tweet Geolocation.* For the second task, one geolocates individual tweets. The approaches of [1, 18] used topic models. Ahmed [1] proposed the nested Chinese Restaurant Franchise Process to derive hierarchical topics whereby topics at a higher level correspond to broad regions and topics at a lower level correspond to more fine-grained locations. Hong et al. [18] employed the Sparse Additive Generative Model framework [11] to model deviations caused by facets, e.g., a posting location coordinate will cause probabilities of certain words in a tweet to "deviate" from some background distribution. Since topics are dependent on the posting coordinates; the topic

models can be used to geolocate tweets by inferring their topics. Priedhorsky et al. [39] modeled each word as a GMM. To geolocate each tweet, the multiple GMMs corresponding to multiple words are linearly combined whereby more LI words are weighted more. The works in [23] used naïve Bayes to model the probability of words given coarse locations such as cities. Given a tweet, one retrieves coarse locations that have a high probability of generating the tweet content. Grid-based approaches [36, 42, 45] have also been explored. Wing and Baldridge [45] discretized space into a uniform grid of square cells, followed by modeling the smoothed distribution of words for each cell. Test tweets are geolocated to the most similar cell based on the Kullback-Leibler (KL) divergence between word distributions or based on tweet content probability under a naïve Bayes model. In [36], O'Hare and Murdock utilize uniform grids, the naïve Bayes language model, and some adaptation of spatial smoothing to geolocate Flickr photos using the photo tags. Instead of uniform grids, the work in [42] proposes an adaptive grid constructed using a k-d tree. This adapts to the training set size and geographic dispersion of the documents, i.e., more densely populated areas will be fitted with more numerous and smaller cells.

For each test tweet, the above works provide either a coordinate estimation [1, 18, 39] or a coarse discrete location, e.g., city/grid cell [23, 36, 42, 45]. Bo Han et al. [17] described this as, respectively, akin to the tasks of multitarget regression and multiclass classification. For the former task, median and mean distance errors are used; for the latter task, classification accuracy is used. In any case, there is a significant difference from fine-grained geolocation, to be discussed next.

### 6.4 Fine-Grained Geolocation

In contrast to coarse-grained geolocation, we work on fine-grained geolocation of tweets. This aims to link tweets to specific venues, e.g., geolocating a tweet "Flight delayed" to some airport venue, instead of a city, grid cell, or a coordinate that may be associated with many venues.

Compared to coarse-grained geolocation, fine-grained geolocation is relatively less explored. However, certain approaches can be carried over. Li et al. [27] modeled each venue as having some distribution over words. In an approach analogous to [45] for coarse-grained geolocation, tweets are geolocated using KL-divergence to the venue with the most similar word distribution. They also model venue probabilities based on posting time. This is linearly combined with the transformed KL-divergences to form venue scores. We implement this approach as a baseline. In [25], each venue generates words according to a fitted naïve Bayes model, analogous to [23] for coarse-grained geolocation. However, not all test tweets will be geolocated. They regard tweets with no LI words as not tractable for geolocation. Such tweets are discarded. Hence, there is a possibility in applications of discarding too many tweets. Ikawa et al. [20] learned the keywords that are highly associated with locations from geocoded tweets generated by location apps. A test tweet that has at least one keyword is then geolocated to the location with highest cosine similarity. Again, there is the issue that test tweets without any key words are ignored. Cao et al. [2] conducted extensive feature engineering with content, location history, and relationships. They used features that are highly specific to Foursquare, e.g., venue categories and user mayorships. The features are used to classify whether a tweet is posted from a venue or not. Our work seeks to develop a more general approach that relies less on platform-specific features. The works by [21, 26] require extracting venue mentions from tweets. Ji et al. [21] proposed a framework to perform location recognition and location linking simultaneously in a joint search space. They formulated fine-grained geolocation as a structured prediction problem and proposed a beam search-based algorithm. Li and Sun [26] extract each location mention in a tweet and predict whether the user has visited, is currently at, or will soon visit the mentioned location. They designed a Conditional Random Field (CRF)–based location tagger, which takes in lexical, grammatical, geographical, and

BILOU[3] schema features. For the discussed works [21, 26], we note that while colloquial mentions are handled, relying on mentions is a bottleneck. For example, a tweet "safely landed" has no mentions, but is indicative of the airport. Mention extraction is also a difficult problem on its own. In our work, we geolocate tweets even if no mentions exist. In fact, manual inspection of a sample of our data shows that venue mentions occur in less than 10% of the tweets.

## 7 CONCLUSION

We show the presence of spatial homophily at fine granularities such that venues near each other are more similar in content. We also show that many users have location history in the form of geocoded tweets and that users are spatially focused, with the tendency to visit venues near each other. Following our empirical studies, we proposed several models for fine-grained geolocation. We achieve large improvements in ranking accuracy with the exploitation of user and venue characteristics, such as user location history and venue temporal popularity.

The negative cases illustrated in Section 5.8.3 also highlight several existing research challenges, which are potential directions for future work. First, users may have sparse or no location history. This applies to tourists, new users (i.e., cold-start problem) or users who simply neglect to geocode any tweets. Second, there exists novelty seeking behavior and/or evolution in posting behavior such that tweets are posted from venues far from one's location history. To handle such scenarios, further work can exploit other characteristics for modeling, e.g., follower–followee relationships in Twitter. We are also interested in content-based collaborative filtering. This is useful in certain scenarios with limited information, e.g., a user may have tweeted frequently but without disclosing location history. If there are other users with similar content and whose location histories exist, then collaborative filtering can be applied to improve geolocation.

## APPENDIX

## A GRADIENTS FOR MODEL NB+S+T+U

Let $\Theta$ represent the model NB+S+T+U. On geolocating a tweet with content $\mathbf{w}$, posting time of day $t$ and posted by user $u$, we have log venue probability as

$$\ln p(v|\mathbf{w}, t, u) \propto \ln p(v|t) + \ln p(u|v) + \ln \sum_{w \in \mathbf{w}} \ln p(w|v). \tag{21}$$

Consider a training tweet with posting venue $v_i$. $v_i$ is paired with non-posting venues in order to contribute to the loss function. For each venue pair $(v, v_i)$ considered in the loss function, we compute

$$\frac{\partial z_\Theta(v_i, v)}{\partial x} = \frac{\partial \ln p(v_i|\mathbf{w}, t, u)}{\partial x} - \frac{\partial \ln p(v|\mathbf{w}, t, u)}{\partial x}, \tag{22}$$

where $x$ represents a model parameter that is to be learnted. Recall from Equation (14) that there are four model parameters: $x_\alpha$, $x_\gamma$, $x_\beta$, and $x_S$. The derivative per training tweet is then

$$\frac{\partial L_\Theta(v_i)}{\partial x} = \sum_{v \neq v_i} \frac{\partial \ln(1 + R_\Theta(v_i, v))}{\partial R_\Theta(v_i, v)} \frac{\partial R_\Theta(v_i, v)}{\partial z_\Theta(v_i, v)} \frac{\partial z_\Theta(v_i, v)}{\partial x}. \tag{23}$$

Summing the derivatives over all training tweets gives the gradient of parameter $x$ with respect to the global loss function. Hence, for each model parameter $x$, we have to compute the derivatives with respect to the log venue probabilities.

[3]BILOU schema identifies the Beginning, Inside, and Last words of a multiword location name, and Unit-length location name.

For the smoothing parameter $x_\alpha$ in $p(w|v)$, we have that

$$\frac{\partial \ln p(v|\mathbf{w},t,u)}{\partial x_\alpha} \propto \frac{\partial \sum_{w\in\mathbf{w}} \ln p(w|v)}{\partial x_\alpha} = \sum_{w\in\mathbf{w}} \frac{2x_\alpha}{c(w,v)+x_\alpha^2+\gamma\psi(v,w)} - \frac{2Wx_\alpha}{c(\cdot,v)+Wx_\alpha^2+\gamma\phi(v)} \tag{24}$$

where $\psi(v,w) = \frac{1}{|nb(v)|}\sum_{v_i\in nb(v)} c(w,v_i)$, $\phi(v) = \frac{1}{|nb(v)|}\sum_{v_i\in nb(v)} c(\cdot,v_i)$ and $\gamma = (1+\exp(-x_\gamma))^{-1}$.

For the weight factor parameter $x_\gamma$ in $p(w|v)$, the derivative is

$$\frac{\partial \ln p(v|\mathbf{w},t,u)}{\partial x_\gamma} \propto \frac{\partial \sum_{w\in\mathbf{w}} \ln p(w|v)}{\partial x_\gamma} = \sum_{w\in\mathbf{w}} \frac{\gamma(1-\gamma)\psi(v,w)}{c(w,v)+x_\alpha^2+\gamma\psi(v,w)} - \frac{\gamma(1-\gamma)\phi(v)}{c(\cdot,v)+Wx_\alpha^2+\gamma\phi(v)}. \tag{25}$$

For the smoothing parameter $x_\beta$ in $p(v|t)$, we compute

$$\frac{\partial \ln p(v|\mathbf{w},t,u)}{\partial x_\beta} \propto \frac{\partial \ln p(v|t)}{\partial x_\beta} = \frac{2x_\beta}{f(v,t)+x_\beta^2} - \frac{2Vx_\beta}{f(\cdot,t)+Vx_\beta^2}. \tag{26}$$

Finally, the derivative for the decay parameter $x_S$ in $p(u|v)$ is as computed in Equation (17).

## B SPATIAL HOMOPHILY EXPERIMENT WITH PARAGRAPH VECTORS

Instead of TFIDF representation, we apply word embedding techniques to represent venues differently. Specifically, we apply *Paragraph Vector* [24]. This extends the word embedding techniques in [31, 32] to learn continuous distributed representation for text segments, such as sentences, paragraphs, or documents. Text segments are represented by tokens, whose embeddings are then learned along with that of words. After embedding, text segments that are semantically similar based on their word context will be close in vector space.

In our context, a text segment is a tweet and the text segment token is the posting venue of the tweet. We embed venues into paragraph vectors such that semantically similar venues (based on their associated tweets) are close in their embedded vectors. For example, on embedding the venues in Singapore (SG-SHT) and using a nightclub venue as the query, the 5 nearest venues in terms of cosine similarities are all nightclub venues.

To ascertain the presence of spatial homophily, we then repeat the steps described in Section 3.1. The main difference is that cosine similarities between venues are now computed using paragraph vectors instead of TFIDF vectors. We use the distributed memory version of paragraph vectors, i.e., PV-DM from [24], and experiment with embedding dimensions (denote as $\epsilon$) of 20 and 40. After training, we again tabulate the proportion of venues whose nearest spatial neighbors are more similar or less similar than sampled non-neighbors. As the embeddings are dense, we did not encounter any venue whose similarities to neighbors are exactly equal to non-neighbors. Also, cosine similarities in this case can be negative and the ratio statistic from Section 3.1 is not meaningful. The average results over 10 runs are tabulated in Table 14.

As can be seen, venues tend to be more similar to their spatial neighbors than non-neighbors. The results are consistent across all categories and datasets, with the "more similar" proportion being larger at 50+%. This trend holds across both embedding dimensions of 20 and 40. In short, even with a different representation of venues, we have arrived at the same conclusion as Section 3.1. Hence, spatial homophily between nearby venues is an established phenomenon, which is easy to uncover empirically.

Table 14. Average Proportion of Venues Where Nearest Neighbors are More (or Less) Similar in Content, Compared to Non-neighbors

| Dataset | Category | More similar ($\epsilon$=20) | Less similar ($\epsilon$=20) | More similar ($\epsilon$=40) | Less similar ($\epsilon$=40) |
|---|---|---|---|---|---|
| SG-SHT | Mixed | 54.99% | 45.01% | 55.52% | 44.48% |
| | Food | 54.15% | 45.85% | 54.58% | 45.42% |
| | Shop | 52.29% | 47.71% | 53.12% | 46.88% |
| SG-TWT | Mixed | 51.68% | 48.32% | 51.67% | 48.33% |
| | Food | 52.65% | 47.35% | 53.42% | 46.58% |
| | Shop | 51.83% | 48.17% | 52.28% | 47.72% |
| JKT-SHT | Mixed | 52.12% | 47.88% | 52.94% | 47.06% |
| | Food | 53.12% | 46.88% | 53.59% | 46.41% |
| | Shop | 56.01% | 43.99% | 56.09% | 43.91% |

*Note*: $\epsilon$ is embedding dimension.

## ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative, and DSO National Laboratories.

## REFERENCES

[1] Amr Ahmed, Liangjie Hong, and Alexander J. Smola. 2013. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. 25–36.

[2] Bokai Cao, Francine Chen, Dhiraj Joshi, and Philip S. Yu. 2015. Inferring crowd-sourced venues for tweets. In *2015 IEEE International Conference on Big Data (Big Data'15)*. 639–648.

[3] Hau-Wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. 2012. @Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM'12)*. IEEE Computer Society, 111–118.

[4] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*. ACM, 759–768.

[5] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*.

[6] Wen-Haw Chong, Bing Tian Dai, and Ee-Peng Lim. 2015. Prediction of venues in Foursquare using flipped topic models. In *Advances in Information Retrieval (ECIR'15). Lecture Notes in Computer Science*, Springer, Berlin.

[7] Wen-Haw Chong and Ee-Peng Lim. 2017. Exploiting contextual information for fine-grained tweet geolocation. In *11th International AAAI Conference on Web and Social Media (ICWSM'17)*.

[8] Konstantina Christakopoulou and Arindam Banerjee. 2015. Collaborative ranking with a push at the top. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*. 205–215.

[9] Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman M. Sadeh. 2012. The Livehoods Project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*.

[10] Thanh-Nam Doan and Ee-Peng Lim. 2016. Attractiveness versus competition: Towards a unified model for user visitation. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM'16)*. 2149–2154.

[11] Jacob Eisenstein, Amr Ahmed, and Eric Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. 1041–1048.

[12] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing Pages (EMNLP'10)*. 1277–1287.

[13] Géraud Le Falher, Aristides Gionis, and Michael Mathioudakis. 2015. Where is the Soho of Rome? Measures and algorithms for finding similar neighborhoods in cities. In *9th International AAAI Conference on Web and Social Media (ICWSM'15)*.

[14] Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. 1977. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software* 3, 3, 209–226.

[15] Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. 2011. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems* 26, 3, 10–14.

[16] Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research* 49, 1, 451–500.

[17] Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter geolocation prediction shared task of the 2016 Workshop on Noisy User-Generated Text. In *Proceedings of the 2nd Workshop on Noisy User-Generated Text (WNUT'16)*. 213–217.

[18] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*. 769–778.

[19] Mans Hulden, Miikka Silfverberg, and Jerid Francom. 2015. Kernel density estimation for text-based geolocation. In *Proceedings of the T29th AAAI Conference on Artificial Intelligence (AAAI'15)*. 145–150.

[20] Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. 2012. Location inference using microblog messages. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12 Companion)*. 687–690.

[21] Zongcheng Ji, Aixing Sun, Gao Cong, and Jialong Han. 2016. Joint recognition and linking of fine-grained locations from tweets. In *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*. 1271–1281.

[22] David Jurgens. 2013. That's what friends are for. Inferring location in online social media platforms based on social relationships. In *7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)*.

[23] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "I'm eating a sandwich in Glasgow": Modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents (SMUC'11)*. 61–68.

[24] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*. 1188–1196.

[25] Kisung Lee, Raghu K. Ganti, Mudhakar Srivatsa, and Ling Liu. 2014. When Twitter meets Foursquare: Tweet location prediction using Foursquare. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous'14)*. 198–207.

[26] Chenliang Li and Aixing Sun. 2014. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14)*. 43–52.

[27] Wen Li, Pavel Serdyukov, Arjen P. de Vries, Carsten Eickhoff, and M. Larson. 2011. The where in the tweet. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*. 2473–2476.

[28] Moshe Lichman and Padhraic Smyth. 2014. Modeling human location data with mixtures of kernel densities. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. 35–44.

[29] Xuelian Long, Lei Jin, and James Joshi. 2013. Understanding venue popularity in Foursquare. In *9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'13)*.

[30] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY.

[31] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv Preprint*1 301.3781.

[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*. 3111–3119.

[33] David M. Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI'08)*. 411–418.

[34] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. 2012. Mining user mobility features for next place prediction in location-based services. In *Proceedings of the IEEE 12th International Conference on Data Mining (ICDM'12)*. 1038–1043.

[35] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. 2011. An empirical study of geographic user activity patterns in Foursquare. In *5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*.

[36] Neil O'Hare and Vanessa Murdock. 2013. Modeling locations with social media. *Information Retrieval* 16, 1, 30–62.
[37] Tatiana Pontes, Gabriel Magno, Marisa Vasconcelos, Aditi Gupta, Jussara Almeida, Ponnurangam Kumaraguru, and Virgilio Almeida. 2012. Beware of what you share: Inferring home location in social networks. In *Proceedings of the IEEE 12th International Conference on Data Mining Workshops (ICDMW'12)*. 571–578.
[38] Tatiana Pontes, Marisa A. Vasconcelos, Jussara M. Almeida, Ponnurangam Kumaraguru, and Virgílio A. F. Almeida. 2012. We know where you live: Privacy characterization of Foursquare behavior. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp'12)*. 898–905.
[39] Reid Priedhorsky, Aron Culotta, and Sara Y. Del Valle. 2014. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'14)*. 1523–1536.
[40] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2015. Twitter user geolocation using a unified text and network prediction model. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*. 630–636.
[41] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. 2011. On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking* 19, 3, 630–643.
[42] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*. 1500–1510.
[43] Dan Tasse, Alex Sciuto, and Jason I. Hong. 2016. Our house, in the middle of our tweets. In *10th International AAAI Conference on Web and Social Media (ICWSM'16)*.
[44] G. Viswanathan, F. Bartumeus, S. V. Buldyrev, J. Catalan, U. Fulco, S. Havlin, M. Da Luz, M. Lyra, E. Raposo, and H. Eugene Stanley. 2002. Levy flight random searches in biological phenomena. *Physica A: Statistical Mechanics and Its Applications* 314, 208–213.
[45] Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11)*. 955–964.
[46] Ning Yang, Xiangnan Kong, Fengjiao Wang, and Philip S. Yu. 2014. When and where: Predicting human movements based on social spatial-temporal events. In *Proceedings of the SIAM International Conference on Data Mining (SDM'14)*. 515–523.
[47] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas S. Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web (WWW'11)*. 247–256.
[48] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*. 186–194.
[49] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat-Thalmann. 2013. Time-aware point-of-interest recommendation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*. 363–372.
[50] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, and Xing Xie. 2014. Mining novelty-seeking trait across heterogeneous domains. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*. 373–384.

Received June 2017; revised September 2017; accepted October 2017